Modellkompetenz im Kontext Biologieunterricht

–

Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biologie
eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I
der Humboldt-Universität zu Berlin
von

Eva Terzer

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I
Prof. Stefan Hecht Ph.D.

Gutachter: 1. Prof. Dr. A. Upmeier zu Belzen

2. Prof. Dr. O. Köller

3. Prof. Dr. H. Prechtl

Tag der mündlichen Prüfung: 19.12.2012

Inhaltsverzeichnis

Abkürzungsverzeichnis

AAAS	*American Association for the Advancement of Science*
bik	Biologie im Kontext
DNA	Desoxyribonukleinsäure
EAP	Expected a Posteriori
FC-Items	*Forced-Choice Items*
ICC	*Item Characteristic Curve*
IRT	*Item Response Theory*
KMK	Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland
MC-Items	Multiple-Choice Items
MSA	Mittlerer Schulabschluss
OECD	*Organisation for Economic Co-operation and Development*
PISA	*Programme for International Student Assessment*
PV	*Plausible Value*
SUMS	*Students' Understanding of Models in Science* (Treagust et al., 2002)
TIMSS	*Third International Mathematics and Science Study*, seit 2003 *Trends in International Mathematics and Science Study*
VOMS	*My Views of Models in Science* (Treagust et al., 2001)
VOSTS	*Views of Science-Technology-Society* (Aikenhead & Ryan, 1992)
wMNSQ	*weighted Mean Square*

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Schülerinnen und Schüler brauchen für einen durch Naturwissenschaften und Technik geprägten Alltag eine naturwissenschaftliche Grundbildung (*scientific literacy*). Damit sie diese im Unterricht entwickeln können, beschreiben die nationalen Bildungsstandards der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) fachspezifische Kompetenzanforderungen, u. a. zum Umgang mit Modellen, für den Mittleren Schulabschluss (MSA). Um Fragen des Kompetenzerwerbs wissenschaftlich fundiert nachgehen und diesen gezielt fördern zu können, ist ein Kompetenzmodell als Referenzsystem notwendig, das detailliert Inhalte und Stufen von Modellkompetenz definiert.

Ein Ansatz zu einer Definition umfasst die deklarative Dimension ‚Kenntnisse über Modelle' mit den Teilkompetenzen ‚Eigenschaften von Modellen' und ‚Alternative Modelle' sowie die prozedurale Dimension ‚Modellbildung' mit den Teilkompetenzen ‚Zweck', ‚Testen' und ‚Ändern von Modellen'. Diese Inhalte werden in drei Niveaus graduiert, die ein unterschiedlich komplexes Wissenschaftsverständnis und unterschiedliche Perspektiven auf Modelle abbilden. Diese Strukturierung und Graduierung von Modellkompetenz wird im hier vorgestellten Projekt für die kognitive Facette mithilfe von Multiple-Choice Items (MC-Items) empirisch überprüft.

Zur Entwicklung eines Tests wurden die Merkmale von MC-Items systematisiert, um den Einfluss nicht-theoriegeleiteter Merkmale zu minimieren und den Bezug zwischen der Kompetenz und ihrer Operationalisierung nachvollziehbar zu machen. Hierzu wurden eine Konstruktionsanleitung ausgearbeitet, Antwort-Möglichkeiten auf der Grundlage von Schülerformulierungen entwickelt und die sprachliche Qualität der MC-Items geprüft.

Um mit diesen Items das Kompetenzmodell empirisch überprüfen zu können, wurde evaluiert, inwiefern die Items das Kompetenzmodell adäquat repräsentieren, inwiefern die Items eine zufriedenstellende psychometrische Qualität haben und inwiefern die Bearbeitung der einzelnen Items als Indikator der

entsprechenden Kompetenz interpretierbar ist. Zunächst wurden mit Blick auf die Repräsentation des Kompetenzmodells durch die Items zwei Experten-Ratings durchgeführt (jeweils $N = 9$). Das erste bezog sich auf die Zuordnung der Itembeschreibungen in der Konstruktionsanleitung zum Kompetenzmodell, das zweite auf die der einzelnen Items selbst. Beide Ratings führten zu einer mindestens guten Übereinstimmung zwischen empirischer und theoretischer Zuordnung (prozentuale Übereinstimmung sowie Cohens κ). Abweichungen betrafen vor allem die Zuordnung der Niveaus, während die Teilkompetenzen fast immer zuverlässig zugeordnet wurden. Darüber hinaus stellte sich die Verbindung von Bereichen des Kompetenzmodells durch den Vergleich von Modell und Original als zentrale Vorstellung heraus.

In einem zweiten Evaluationsschritt wurde die psychometrische Qualität der Items geprüft. Hierzu bearbeiteten Schülerinnen und Schüler in vier Teilstudien 191 MC-Items ($N = 1229$). Nach Itemschwierigkeit, Trennschärfe, *Itemfit* sowie *Item Characteristic Curve* (ICC) wurden 45 MC-Items selektiert, die das Spektrum der Personenfähigkeiten insgesamt sinnvoll abdecken. Eine eher geringe Varianz in der Modellkompetenz der Schülerinnen und Schüler ging mit einer entsprechend geringen Reliabilität einher. Da der Test nicht auf eine Individualdiagnose, sondern auf die Prüfung einer Struktur auf Populationsebene abzielt, ist seine Verwendung für die empirische Strukturierung und Beschreibung von Modellkompetenz trotz der geringen Reliabilität vertretbar.

Das dritte Evaluationskriterium bezieht sich auf die Interpretierbarkeit der Items als Indikatoren für Modellkompetenz. Inwiefern während der Itembearbeitung kognitive Prozesse ablaufen, die im Sinne von Modellkompetenz interpretiert werden können, wurde mit der Methode des lauten Denkens erhoben. Diese Denkprotokolle wurden nach einem Codierleitfaden inhaltsanalytisch ausgewertet. Die Items wurden dann als valide eingestuft, wenn 1. Vorstellungen in der angesteuerten Teilkompetenz und Niveaustufe formuliert wurden und 2. im Sinne von Modellkompetenz fachlich angemessene Überlegungen mit einer richtigen Beantwortung bzw. nicht angemessene mit einer falschen Beantwortung verbunden waren. Nach diesen Kriterien konnten 40 von 45

Items als valide eingestuft werden. Der Vergleich von Modell und Original zeigte sich auch hier als prominente Vorstellung.

Für die empirische Beschreibung von Modellkompetenz wurden 40 MC-Items sowie Fragebögen zu weiteren Konstrukten in einem Multi-Matrix-Design Schülerinnen und Schülern des Gymnasiums in den Jahrgangsstufen 7 bis 10 vorgelegt (N = 1136). Die Daten wurden mit ein-, zwei- und fünfdimensionalen IRT-Modellen analysiert, um die Struktur von Modellkompetenz empirisch zu überprüfen. Die Itemschwierigkeiten des Modells, das die beste Passung mit den Daten aufweist, wurden zur empirischen Überprüfung der Graduierung von Modellkompetenz in einer einfaktoriellen Varianzanalyse genutzt. Die Personenfähigkeiten, die im Modell mit der besten Passung geschätzt wurden, gingen in zwei weitere Analysen ein. Zum einen wurde in einer latenten Regression geprüft, inwieweit sich die Jahrgangsstufen 7 bis 10 in ihrer Modellkompetenz unterscheiden. Zum anderen wurden Beziehungen zwischen Modellkompetenz und anderen Variablen korrelationsanalytisch untersucht.

Das eindimensionale Modell wies die beste Passung mit den Daten auf, was für eine eindimensionale Struktur von Modellkompetenz spricht. Messfehlerkorrigierte Korrelationen in mittlerer Höhe deuten möglicherweise auf spezifische Verknüpfungen von Teilkompetenzen hin. Die Graduierung in drei Niveaus erklärte substanzielle Anteile der Varianz und die Niveaus unterscheiden sich statistisch signifikant voneinander. Der Kompetenzunterschied von der Jahrgangsstufe 7 bis 10 war ebenfalls signifikant und band erhebliche Varianzanteile. Das Alter spielte hierfür keine Rolle. Personenfähigkeiten im Bereich Modellkompetenz hingen mit Schulnoten in naturwissenschaftlichen wie sprachlichen Fächern in vergleichbarer Höhe zusammen ($-.233 < M_r < -.369$). Von allgemeinen kognitiven Fähigkeiten und Lesefähigkeiten war Modellkompetenz mit $M_{r\,KFT} = .351$ bzw. $M_{r\,LGVT\_G} = .523$ und $M_{r\,LGVT\_V} = .486$ abgrenzbar. Ein Unterschied zwischen den Geschlechtern zeigte sich nicht.

Auch wenn es Hinweise auf eine eindimensionale empirische Struktur von Modellkompetenz gibt, sprechen die Befunde dafür, dass die Teilkompetenzen spezifisch miteinander verknüpft sind. Die Niveaus bilden insgesamt steigende

Anforderungen an die Schülerinnen und Schüler ab. Lehrerinnen und Lehrer können das Kompetenzmodell somit sinnvoll als Referenzsystem heranziehen. Es zeichnet sich ab, dass eine domänen- bzw. kontextspezifische und eine übergreifende Facette von Modellkompetenz differenziert werden können. Diese sollten gezielt gefördert werden. Modellkompetenz ist nach den hier vorgestellten Befunden grundsätzlich erlernbar. Schülerinnen und Schüler in der Jahrgangsstufe 7 verfügen vor allem über Kompetenzen im medialen Bereich, bis Jahrgangsstufe 10 verschiebt sich der Schwerpunkt hin zu einer methodischen Perspektive auf Modelle. Während sich Modellkompetenz in einer diskriminanten Validierung von allgemeinen kognitiven Fähigkeiten und Lesefähigkeiten sinnvoll abgrenzen lässt, bleibt das Bild für die konvergente Validierung noch unklar.

Für die Modellierung von Zusammenhängen zwischen Niveaus einzelner Teilkompetenzen, für andere Anwendungskontexte wie z. B. ein individuelles Feedback, für andere Zielgruppen wie z. B. Studierende sowie für die weitere Untersuchung schwierigkeitserzeugender Itemmerkmale, z. B. den Kontext, sollte der Itempool für anschließende Studien erweitert bzw. adaptiert werden. Weitere Anschlussfragestellungen beziehen sich auf die Kontextspezifität von Modellkompetenz und die Gestaltung von entsprechenden Unterrichtsangeboten.

1 Einleitung

Ob ein Zahntechniker für den Zahnersatz ein Modell herstellt oder eine Bürgerin sich über Klimabilanzen informieren möchte – Modelle begegnen dem Menschen in einem durch die Naturwissenschaften und Technik geprägten Alltag überall. Sowohl zur Vorbereitung auf das Berufsleben als auch für die gesellschaftliche Teilhabe ist es deshalb wichtig, Kompetenzen im Umgang mit Modellen zu erlangen - die Frage ist, wie Schülerinnen und Schüler diese im Biologieunterricht erwerben können. Welche Inhalte sind mit Blick auf Modelle in der Biologie relevant, um mit ihnen reflektiert umgehen zu können? Gibt es unterschiedlich komplexe Vorstellungen zu diesen Inhalten? Können diese Vorstellungen aufeinander aufbauend oder in einer beliebigen Reihenfolge erworben werden? Welche Aufgaben sind geeignet, um bereits verfügbare Kompetenzen zu erfassen bzw. Kompetenzen zu erwerben? Antworten auf diese Fragen geben Kompetenzmodelle, die durch eine Beschreibung von Kompetenzinhalten und -stufen sowie entsprechende fachbezogene Aufgaben und Testverfahren konkrete Anhaltspunkte für eine gezielte Förderung bieten. Somit dienen Kompetenzmodelle als Referenzsystem für das professionelle Handeln von Lehrkräften (Klieme, Avenarius et al., 2007). Der Begriff der Kompetenz beschreibt dabei die kontextspezifische „Verbindung von Wissen und Können in der Bewältigung von Handlungsanforderungen“ (Klieme & Hartig, 2007, S. 19), die in kognitive und motivationale Facetten differenziert werden kann (Klieme & Leutner, 2006; Kapitel 2.1).

Die empirische Bildungsforschung nimmt Kompetenzen als *Outcome* von Bildungsprozessen in den letzten zehn Jahren stärker in den Blick als den *Input* in das Bildungssystem durch Steuerungsdokumente wie z. B. Rahmenlehrpläne (z. B. Bybee, 2002; Klieme, Maag-Merki & Hartig, 2007; Köller, 2008a; Koeppen, Hartig, Klieme & Leutner, 2008). Ausschlaggebend für diesen Para-

digmenwechsel waren die Ergebnisse von TIMSS[1] und PISA[2]: Schülerinnen und Schüler erreichten häufig nicht die in den Rahmenlehrplänen der Bundesländer formulierten Ziele (Klieme & Hartig, 2007; Klieme & Steinert, 2004).

Ziele im Bereich der Naturwissenschaften beschreibt das Konzept der *scientific literacy*. Darin wird formuliert, welche Kompetenzen Schülerinnen und Schüler in diesem Bereich als berufliche Qualifikation sowie für die gesellschaftliche Teilhabe benötigen (Baumert, 1997; Bybee, 1997, 2002). Diese Kompetenzen sollten im Sinne eines lebenslangen Lernens „für die weitere schulische und berufliche Entwicklung von Bedeutung [sein] und anschlussfähiges Lernen fördern" (Rost, Prenzel, Carstensen, Senkbeil & Groß, 2004, S. 123). In Anlehnung an das *Literacy*-Konzept von PISA legen die deutschen Bildungsstandards nun fachspezifische Kompetenzanforderungen fest[3], denen Schülerinnen und Schüler mit dem Mittleren Schulabschluss in der Regel genügen sollen (KMK, 2005). Sie konkretisieren somit abstrakte Bildungsziele und folgen dem internationalen Trend, Entwicklungspunkte für Schülerinnen und Schüler in Standards bzw. *Benchmarks* zu definieren (z. B. *American Association for the Advancement of Science* (AAAS), 1993; *National Research Council*, 1996). Dabei beinhaltet die Orientierung an Standards die oben angesprochene zielbezogene Orientierung an Kompetenzmodellen.

---

[1] *Third International Mathematics and Science Study*, seit 2003 *Trends in International Mathematics and Science Study*.

[2] *Programme for International Student Assessment*.

[3] „Das Grundbildungskonzept von PISA und die PISA-Tests sind in der Diskussion in Deutschland weitgehend als Maßstäbe der Bildungsqualität akzeptiert worden, sodass es sinnvoll erscheint, auch die nationalen Standards in Beziehung zu PISA zu setzen und an denselben Schülergruppen (also im 9. Jahrgang) zu untersuchen" (Klieme, Avenarius et al., 2007, S. 138).

In internationalen Steuerungsdokumenten schlagen sich Modelle als wichtiges Produkt von Wissenschaft sowie als wissenschaftliche Methode nieder (Boulter & Gilbert, 2000; Driver, Leach, Millar & Scott, 1996). In den Bildungsstandards der KMK (2005) wird als Ziel formuliert, dass Schülerinnen und Schüler Modelle zum Bearbeiten, Veranschaulichen, Erklären und Beurteilen komplexer Phänomene nutzen (E9-E12) sowie kritisch über Modelle reflektieren (E13; KMK, 2005). Auch wenn in den Bildungsstandards die Bedeutung von Modellen als wissenschaftliche Methode eher im Hintergrund steht, wird der Umgang mit Modellen im Kompetenzbereich Erkenntnisgewinnung verortet. Dieser erhält durch den engen Bezug der deutschen Bildungsstandards zum *Literacy*-Konzept eine bedeutende Rolle. Dies geht mit einer Verlagerung der Aufmerksamkeit vom fachlichen Inhalt eines Modells zu seinen Charakteristika und Funktionen einher (vgl. Van Driel & Verloop, 1999).

Die Standards bilden die wesentliche Funktion von Modellen in der Wissenschaft ab, fokussiert Merkmale der Welt zu rekonstruieren. Diese Rekonstruktion ermöglicht es, Dinge zu visualisieren, zu kommunizieren sowie hypothetisch-deduktiv Erkenntnisse über die Welt zu gewinnen (Giere, 2004; Stachowiak, 1973). Da die Erkenntnisgewinnung und Theoriebildung insbesondere in den Naturwissenschaften häufig mit Modellen verbunden sind (Oh & Oh, 2011), wird Wissenschaft in einem *model-based view of scientific theory and scientific inquiry* sogar als Konstruktion von Modellen gesehen (Giere, 1988; Gilbert, 1991; Van Fraassen, 1980). Modellkompetenz trägt sowohl zu einem tieferen Verständnis von Fachinhalten bei als auch zu einem elaborierten Wissenschaftsverständnis und Kompetenzen im wissenschaftlichen Denken (Gobert & Pallant, 2004; Lehrer & Schauble, 2006; Schwarz & White, 2005; Sins, Savelsbergh, van Joolingen & van Hout-Wolters, 2009).

Diese potenziellen positiven Effekte von Modellkompetenz werden jedoch nicht ausgeschöpft: Nationale und internationale Untersuchungen zeigen, dass Schülerinnen und Schüler Modelle kaum als Werkzeuge der wissenschaftlichen Erkenntnisgewinnung betrachten und wenig reflektiert mit Modellen umgehen (z. B. Grosslight, Unger, Jay & Smith, 1991; Treagust, Chittleborough & Mamiala, 2002; Trier & Upmeier zu Belzen, 2009). Auch in Problemlösesituationen,

die das Heranziehen von Modellen erfordern, haben Schülerinnen und Schüler laut PISA 2003 (Prenzel et al., 2004) Schwierigkeiten. Die verfügbaren Kompetenzen der Schülerinnen und Schüler erfüllen demnach nicht den Anspruch, der in den Bildungsstandards formuliert wird. Somit werden im Bereich Modellkompetenz weder entsprechende berufsbezogene Qualifikationen erreicht noch die Partizipation am gesellschaftlichen Diskurs vorbereitet. Hierfür ist eine effektive Förderung im Bereich Modellkompetenz notwendig.

Welche Schritte beim Erwerb von Modellkompetenz relevant sind und an welchen Punkten eine Förderung anknüpfen kann, ist bislang nicht ausreichend untersucht. Die Grundlage hierfür bildet ein theoretisch hergeleitetes Kompetenzmodell der Modellkompetenz von Upmeier zu Belzen und Krüger (2010). Dieses Modell ermöglicht es, Ziele für den Umgang mit Modellen zu konkretisieren, Diagnosen einzuordnen und gezielte Fördermaßnahmen zur Entwicklung von Modellkompetenz im Biologieunterricht zu entwickeln. Dies kann jedoch nur dann gelingen, wenn es die Aspekte der Kompetenz von Lernenden, ihre Niveaustufung sowie ihre Entwicklung widerspiegelt (Klieme, Avenarius et al., 2007). Um dies empirisch prüfen zu können, wird das Kompetenzmodell in verschiedenen Aufgabenformaten operationalisiert (Grünkorn & Krüger, 2012; Krell & Krüger, 2010; Terzer, Hartig & Upmeier, angen.). Ziel der vorliegenden Untersuchung ist es, die kognitive Facette des Kompetenzmodells bezogen auf den konkreten Umgang mit Modellen in Items zu operationalisieren, diese Operationalisierng zu evaluieren sowie Modellkompetenz empirisch zu beschreiben. Dies macht eine gezielte Förderung der Schülerinnen und Schüler im Bereich Modellkompetenz überhaupt erst möglich, um die Lücke zwischen Bildungsanspruch und tatsächlicher Kompetenz zu schließen. Zusätzlich stellt das vorliegende Projekt ein Diagnoseinstrument bereit, das z. B. zur Evaluation von Interventionsstudien auf Klassenebene genutzt werden kann.

2 Theorie

Um den Kompetenzerwerb von Schülerinnen und Schülern und damit Bildungsziele im Bereich Modellkompetenz gezielt unterstützen zu können, wird ein entsprechendes Kompetenzmodell benötigt (Kapitel 1). Dieses muss für einen effektiven Einsatz zunächst empirisch überprüft werden (Hartig & Klieme, 2006; Klieme, Avenarius et al., 2007). Dies ist das Kernziel der vorliegenden Arbeit.

Wesentliche theoretische Grundlage dieser Arbeit ist somit der Kompetenzbegriff (z. B. Klieme & Hartig, 2007; Weinert, 2001; Kapitel 2.1). Hierzu werden zwei Ebenen unterschieden (Abb. 1): eine latente Ebene mit dem Konstrukt Modellkompetenz und eine manifeste Ebene der Performanz im Bereich Modellkompetenz, d. h. als gezeigtes Verhalten. Im Gegensatz zum Konstrukt selbst ist die Performanz direkt beobachtbar. Für die empirische Überprüfung der Beschreibung des Konstrukts Modellkompetenz wird die Performanz mit einem Test erhoben.

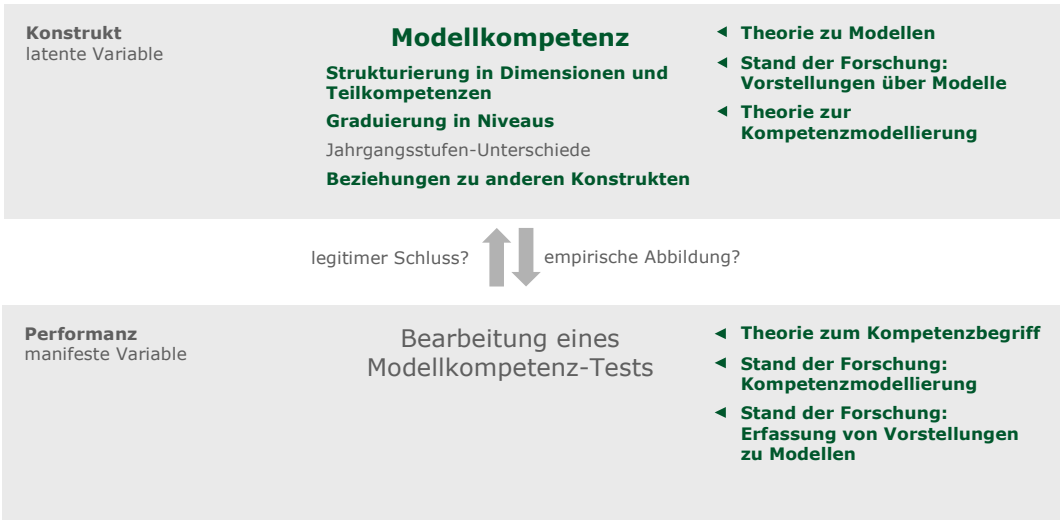


Abb. 1: Konzeption des Projekts – Einbindung der theoretischen Grundlagen.

Der Stand der Forschung zur Modellierung von Kompetenzen bietet eine Orientierung für die Entwicklung eines Vorgehens zu empirischen Beschreibung

von Modellkompetenz (Abb. 1). In der Didaktik der Biologie war das Projekt Biologie im Kontext (bik; Bayrhuber et al., 2007) der Auftakt zur Kompetenzmodellierung und wird deshalb hierzu herangezogen.

Um die Modellierung einer Kompetenz empirisch überprüfen zu können, muss diese in Theorien der jeweiligen Domäne verankert werden (Köller, 2008b). Für Modellkompetenz im Bereich der Biologie sind dies wesentliche wissenschaftliche Konzepte zu Modellen und ihrer Anwendung (z. B. Giere, 2006; Johnson-Laird, 1983; Mahr, 2008a; Stachowiak, 1973; Kapitel 2.2) sowie empirische Untersuchungen zu entsprechenden Vorstellungen, über die Schülerinnen und Schüler verfügen (z. B. Grosslight et al., 1991; Kapitel 2.3; Abb. 1).

Auf diesen Grundlagen beruht das Kompetenzmodell der Modellkompetenz von Upmeier zu Belzen und Krüger (2010), das dieser Arbeit zugrunde gelegt wird (Kapitel 2.4). Die Domänenspezifität von Modellkompetenz wird näher beschrieben, indem sie in ein nomologisches Netzwerk eingebettet wird, d. h. Beziehungen von Modellkompetenz zu anderen Konstrukten dargestellt werden (Köller, 2008b; Abb. 1).

2.1 Kompetenz

Der Kompetenzbegriff wurde Anfang der 1970er Jahre über die Linguistik (Chomsky, 1968) und die pragmatisch-funktionale Tradition der amerikanischen Psychologie (McClelland, 1973; White, 1959) in die Sozialwissenschaften eingeführt. Seitdem verlagert sich der Fokus von traditionellen Konzepten der allgemeinen Bildung zunehmend auf die Fähigkeit, Anforderungen in realen Situationen, wie z. B. im Berufsleben, zu bewältigen[4]. Dieser neue Fokus soll eine bessere Vorhersage von Leistungsunterschieden in berufsbezogenen

[4] Für eine ausführliche Darstellung der Geschichte des Kompetenzbegriffs siehe Vonken (2005).

Qualifikationen ermöglichen. Dazu werden Inhalte von prognostischen Tests enger mit dem *Outcome*, d. h. diesen Qualifikationen, verbunden (Klieme & Hartig, 2007). Das Augenmerk liegt somit auf der Handlungsfähigkeit von Personen in unterschiedlichen Kontexten (z. B. Connell, Sheridan & Gardner, 2003, S. 131: Kompetenz als „*realised abilities*"; White, 1959, S. 318: Kompetenz als „*effective interaction … [of the individual] with the environment*"), die im Sinne einer *scientific literacy* die Partizipation an gesellschaftlichen Diskussionen sowie das lebenslange Lernen ermöglichen soll (vgl. Kapitel 1). Entsprechend sind der Kompetenzbegriff und Theorie zu Kompetenzmodellen Grundlage dieser Arbeit (Kapitel 2.1.1). Die Kompetenzforschung befasst sich mit der Entwicklung 1. von Kompetenzmodellen, 2. entsprechender psychometrischer Modelle und 3. Messverfahren sowie 4. der Interpretation von Kompetenzmessungen für Akteurinnen und Akteure im Bildungswesen (Klieme & Leutner, 2006; Koeppen et al., 2008). Der Stand der Forschung in der Didaktik der Biologie, hier exemplarisch für das erste große Kompetenzmodellierungsprojekt bik (Bayrhuber et al., 2007) vorgestellt, bezieht sich bislang vor allem auf die ersten drei Forschungsbereiche (Kapitel 2.1.2).

2.1.1 Kompetenzbegriff und Kompetenzmodelle

In den Bildungsstandards der KMK (2005) werden *can do statements* formuliert, z. B. „Die Schülerinnen und Schüler wenden Modelle zur Veranschaulichung von Struktur und Funktion an" (E9, KMK, 2005, S. 14). Somit wird Kompetenz dort als Performanz verstanden und nicht zwischen einer latenten und einer manifesten Ebene unterschieden. Dagegen verwendet die empirische Bildungsforschung einen stärker psychologisch geprägten Kompetenzbegriff nach Weinert (2001; enger gefasst in Klieme & Leutner, 2006) und trennt diese Ebenen. Hier wird Kompetenz als verfügbare Fähigkeiten und Fertigkeiten verstanden, die der Performanz zugrunde liegen. Der Kompetenzbegriff bezieht sich damit in der empirischen Bildungsforschung auf ein Konstrukt, das im Gegensatz zur Performanz nicht direkt beobachtbar ist und deshalb operationalisiert werden muss (Köller, 2008a).

Dieser Kompetenzbegriff ist funktional geprägt – Klieme und Hartig (2007, S. 19) definieren Kompetenz etwa als „Verbindung von Wissen und Können in

der Bewältigung von Handlungsanforderungen". Dabei ist die Frage zentral, *wofür* jemand kompetent ist bzw. sein soll. Durch diese Kontextspezifität sowie durch die Erlern- und Vermittelbarkeit, die sich daraus ergibt, ist Kompetenz von situationsunabhängigen allgemeinen kognitiven Leistungsdispositionen, d. h. Intelligenz, abzugrenzen (Hartig & Klieme, 2006; Klieme & Hartig, 2007; Klieme & Leutner, 2006; Klieme, Maag-Merki et al., 2007; Koeppen et al., 2008; McClelland, 1973).

Um die Struktur eines Kompetenzmodells zu entwickeln, werden entsprechend relevante Situationen und Anforderungen herangezogen und die „mentalen Bedingungen" (Klieme & Hartig, 2007, S. 19) rekonstruiert, die das Handeln beeinflussen. In einer Situation betrifft dies neben kognitiven Komponenten auch motivationale, volitionale und soziale Bereitschaften und Fähigkeiten wie z. B. Einstellungen, Komponenten der Selbstregulation oder motivationale Orientierungen (Klieme & Hartig, 2007; Weinert, 2001). Indem die kognitiven Komponenten getrennt von anderen erfasst werden, werden die Zusammenhänge zwischen verschiedenen spezifischen kognitiven Kompetenzen, nichtkognitiven Variablen wie z. B. Einstellungen sowie allgemeinen kognitiven Fähigkeiten empirisch zugänglich. Dies ermöglicht die Untersuchung von Effekten unterschiedlicher individueller Lernvoraussetzungen und Entwicklungsprozesse (Klieme & Leutner, 2006; Klieme, Maag-Merki et al., 2007).

Für die empirische Untersuchung von Kompetenzen ist es notwendig, individuelle Kompetenzausprägungen möglichst eindeutig bestimmen zu können. Aufbauend auf die Definition eines spezifischen Kompetenzkonstrukts bedarf es eines Tests, der diese Ausprägungen misst (Klieme & Hartig, 2007). Die Grundlage für solche Tests bilden Kompetenzmodelle, die das jeweilige spezifische Konstrukt strukturiert beschreiben.

Bei der Entwicklung eines Kompetenzmodells sollte eine Spezifizierung und Operationalisierung der Kompetenz zunächst in den Domänen bzw. Fächern erfolgen und auf psychologischen und insbesondere fachdidaktischen Theorien aufbauen (Klieme, Avenarius et al., 2007). Klieme, Avenarius et al. (2007, S. 75) heben hier insbesondere die Bedeutung der Fachdidaktik hervor, da

diese „Lernprozesse in ihrer fachlichen Systematik und zugleich in der je spezifischen, domänen-abhängigen Logik des Wissenserwerbs und der Kompetenzentwicklung" rekonstruiert und somit der Kontextspezifität von Kompetenz Rechnung trägt. Dies sollte für die Entwicklung von Kompetenzmodellen berücksichtigt werden.

Die Kompetenzmodellierung kann mit unterschiedlichen Zielen verbunden sein, die die Entwicklung eines Kompetenzmodells leiten: Sie kann 1. auf ein individuelles Feedback bzw. individuelle Laufbahnentscheidungen abzielen, 2. auf die summative Evaluation von Lernerfolg auf aggregierter Ebene zur Bilanzierung von Bildungsprozessen (Lernstandserhebungen bis zu *large-scale assessments* im Rahmen von Systemmonitoring, z. B. bei TIMSS und PISA) oder 3. auf Grundlagenforschung wie z. B. die Beschreibung von Kompetenzentwicklung oder die Beurteilung von Interventionen (Klieme, Avenarius et al., 2007; Koeppen et al., 2008; Leutner, Fleischer, Spoden & Wirth, 2007). Ein Kompetenzmodell sollte eindeutig einem solchen spezifischen Zweck zugeordnet werden, da z. B. eines für individuelles Feedback deutlich detaillierter sein muss als eines, das zur vergleichenden Evaluation von Bildungsinstitutionen herangezogen werden soll (Klieme & Leutner, 2006; Koeppen et al., 2008).

Die Kompetenzmodellierung umfasst vier Forschungsbereiche (Klieme & Leutner, 2006; Klieme, Hartig & Rauch, 2008; Koeppen et al., 2008):

1. Modellierung von Kompetenzen

Zunächst muss theoriegeleitet ein Kompetenzmodell formuliert werden. Die Definition des Kontextes, d. h. auf welche Situationen und Anforderungen sich ein spezifisches Kompetenzkonstrukt bezieht, ist zentral für eine sinnvolle Fassung des jeweiligen Konstrukts und darf weder zu abstrakt noch zu eng sein (Klieme, Maag-Merki et al., 2007).

Wie die Bewältigung unterschiedlicher Anforderungen in entsprechenden Situationen miteinander zusammenhängt, wird – zunächst hypothetisch – in einem *Strukturmodell* beschrieben, das eine Kompetenz in Teilkompetenzen strukturiert. Unterschiedliche Ausprägungen spezifischer Fähigkeiten werden

in *Niveaumodellen* definiert, die, wenn möglich a priori, eine Graduierung in Niveaus beschreiben (Hartig & Klieme, 2006). Dies ermöglicht es, Testergebnisse kriteriumsorientiert, d. h. mit Bezug auf ein vorab festgelegtes Kriterium, zu interpretieren (Klieme, Avenarius et al., 2007). Die Niveaus können außerdem als kontinuierliche Entwicklung oder als qualitative Sprünge interpretiert werden (Klieme & Leutner, 2006; Koeppen et al., 2008). Inwiefern dies der Fall ist und hierarchische Niveaus einer Kompetenz tatsächlich zeitlich aufeinander folgen, aufeinander aufbauen und somit über ein Niveaumodell hinaus ein *Kompetenzentwicklungsmodell* bilden, ist empirisch zu prüfen (Schecker & Parchmann, 2006). Ein Kompetenzmodell kann entsprechend sowohl ein Struktur- als auch ein Niveau- und Entwicklungsmodell sein: Die Kategorisierung als Struktur-, Niveau- oder Entwicklungsmodell schließt sich nicht gegenseitig aus, sondern ergänzt sich. Ein Kompetenzmodell bietet durch die strukturierte Beschreibung einer Kompetenz Kriterien, die Passung von Bildungsangeboten zu intendierten Zielen zu evaluieren (Hartig & Klieme, 2006; Klieme, Maag-Merki et al., 2007).

Schecker und Parchmann (2006) differenzieren neben dieser Kategorisierung *normative Kompetenzmodelle*, die theoriegeleitet oder historisch gewachsen zu erreichende Kompetenzen beschreiben, und *deskriptive Kompetenzmodelle*, die typische Muster erworbener Kompetenz abbilden. Sie plädieren dafür, normative Modelle empirisch zu prüfen. Beispiele hierfür sind etwa die länderübergreifenden Bildungsstandards (KMK, 2005) mit vier Kompetenzbereichen als Dimensionen, die in jeweils drei Anforderungsbereiche graduiert sowie in drei inhaltliche Basiskonzepte differenziert sind, oder Rahmenlehrpläne, denen im Sinne eines Spiralcurriculums implizit ebenfalls ein Kompetenzentwicklungsmodell zugrunde liegt (z. B. Senatsverwaltung für Bildung, Jugend und Sport Berlin, 2006).

2. Entwicklung entsprechender psychometrischer Modelle

Die Kompetenzmodellierung erfordert anspruchsvolle psychometrische Methoden und multidisziplinäre Forschungsaktivitäten (Koeppen et al., 2008). Bei der Entwicklung psychometrischer Modelle stellt sich die Frage, „(…) auf welche Weise das Verhältnis zwischen situativen Anforderungen einerseits und Personenmerkmalen andererseits modelliert werden muss, damit individuelle Testwerte als Fähigkeit zur Bewältigung spezifischer Anforderungen interpretiert werden können" (Klieme & Leutner, 2006, S. 8). Ein psychometrisches Modell vermittelt somit zwischen einem Testergebnis und einem Kompetenzmodell: Es begründet, wie ein Ergebnis im Sinne des Kompetenzmodells zu interpretieren ist, und definiert die angenommene Beziehung zwischen der latenten und manifesten Ebene, d. h. zwischen den benötigten Kompetenzen in einer spezifischen Domäne und der Bearbeitung eines Kompetenztests (Klieme & Leutner, 2006). Beispiele hierfür sind das Rasch-Modell oder multidimensionale[5] IRT-Modelle (Kapitel 4.2.1), die die Itemschwierigkeit und die Personenfähigkeit zueinander in Beziehung setzen (Hartig, 2008). Wie differenziert dabei Teilkompetenzen in einem psychometrischen Modell erfasst werden, wird unter Abwägung ökonomischer und theoretischer Argumente entschieden (Klieme, Maag-Merki et al., 2007).

---

[5] Der Begriff „Dimension" wird im Kontext psychometrischer Modelle anders verwendet als auf Ebene von Kompetenzmodellen. Die Begriffe „Teilkompetenz" und „Dimension" werden hier mit Blick auf Kompetenzmodelle so verwendet, dass Teilkompetenzen in Dimensionen gruppiert sind (z. B. Dimension ‚Kenntnisse über Modelle' mit den Teilkompetenzen ‚Eigenschaften von Modellen' und ‚Alternative Modelle', Kapitel 2.4.1). Auf der Ebene von psychometrischen Modellen werden Dimensionen als zusammenhängende Itemgruppen verstanden, d. h. sowohl Teilkompetenzen als auch Dimensionen von Kompetenzmodellen bilden jeweils Dimensionen psychometrischer Modelle. Der Begriff „Facette" bezieht sich auf verschiedene Anteile von Kompetenz, z. B. kognitive und affektive oder domänenspezifische und domänenübergreifende.

3. Test- und Itemkonstruktion auf der Grundlage von Kompetenzmodellen

Da Kompetenz kontextspezifisch ist, müssen Items zu ihrer Messung realen Situationen möglichst ähnlich sein (Hartig & Klieme, 2006). Insbesondere mit Blick auf den funktional definierten Kompetenzbegriff (vgl. Kapitel 2.1.1) ist deshalb die Validität von Kompetenztests zu prüfen. Eine weitere Anforderung an Kompetenzitems ist, sie eindeutig einem Bereich des jeweiligen Kompetenzmodells zuzuordnen. Das bedeutet, dass sie so eng formuliert und so klar von Items zu anderen Bereichen des Kompetenzmodells abgegrenzt sein müssen, dass eine solche Zuordnung eindeutig ist. Inwiefern dies gelungen ist, können Expertenratings empirisch klären (Schecker & Parchmann, 2006).

4. Nutzung der diagnostischen Information aus Kompetenztests

Je nachdem, worauf der Einsatz eines Kompetenztests abzielt (s. o.), muss die diagnostische Information für unterschiedliche Akteurinnen und Akteure im Bildungswesen unterschiedlich aufbereitet werden. Ein individuelles Feedback in einer Domäne sollte z. B. so detailliert sein, dass es als Grundlage für eine gezielte Kompetenzentwicklung dienen kann. Beim Bildungsmonitoring werden dagegen breit mehrere Domänen betrachtet, um die Kompetenzverteilung in Bundesländern oder Staaten abzubilden und Steuerungsentscheidungen vorzubereiten. Die diagnostische Information muss entsprechend unterschiedlich grob aufgelöst aufbereitet werden und stellt z. B. unterschiedliche Ansprüche an die Reliabilität eines Tests (Leutner et al., 2007).

2.1.2 Kompetenzmodellierung in der Didaktik der Biologie

Im Bereich der Naturwissenschaften gibt es ein breit angelegtes, eher grob aufgelöstes Niveaumodell von Bybee (2002). Es unterscheidet vier Niveaus von *scientific literacy*: nominal (Zuordnung von Begriffen und Fragen zum Bereich Naturwissenschaft), funktional (Kenntnis von Fachbegriffen), konzeptionell und prozedural (Verständnis von Konzepten und Prozessen, Struktur von Disziplinen) sowie multidimensional (umfassendes Wissenschaftsver-

ständnis). In Anlehnung daran wurde für naturwissenschaftliche Kompetenzen in PISA 2006 (Prenzel, Carstensen, Frey, Drechsel & Rönnebeck, 2007; Prenzel, Schöps et al., 2007) ein Kompetenzniveaumodell formuliert. Es zielt darauf ab, im Sinne einer summativen Evaluation einen breit gefächerten Überblick über den Status quo des deutschen Bildungssystems zu geben. Das Kompetenzstrukturmodell von PISA 2006 ist in vier Inhaltsbereiche, drei Prozesse als Teilkompetenzen und fünf Typen von Situationen gegliedert und bezieht sich vor allem auf fachübergreifende naturwissenschaftliche Kompetenzen (Prenzel, Carstensen et al., 2007). Es ersetzt aufgrund der Kontextspezifität von Kompetenz keine Kompetenzmodelle für einzelne Domänen.

Im Fach Biologie wurde von 2005 bis 2008 auf Grundlage der Bildungsstandards der KMK (2005) das Projekt bik durchgeführt. Es zielte darauf ab, Kompetenzen der Schülerinnen und Schüler in den vier Kompetenzbereichen ‚Fachwissen', ‚Erkenntnisgewinnung', ‚Kommunikation' und ‚Bewertung' zu fördern und innovative Unterrichtskonzeptionen sowie Aufgaben für den Biologieunterricht zu entwickeln, zu evaluieren und zu implementieren (Bayrhuber et al., 2007). Dieses Pilotprojekt zeichnet sich dadurch aus, dass hier erstmals Ansätze zur empirischen Überprüfung von Kompetenzmodellen im Bereich der Didaktik der Biologie entwickelt wurden. Diese werden als Grundlage für einen hier zu entwickelnden Ansatz aufgegriffen.

Die Herangehensweise der bik-Teilprojekte unterschied sich je nach Art der untersuchten Kompetenz. Für Kompetenzen, die sich stark auf Denkprozesse beziehen, wurden eher qualitative Zugänge mit Interviews oder Items in offenem Antwortformat genutzt (Eggert, 2008; Kramer, 2009; Reitschert & Hößle, 2007), während konzeptuelles Wissen mit Items in geschlossenem Antwortformat erhoben wurde (Grube, 2011; Schmiemann, 2010). Die jeweiligen verwendeten Items wurden vor ihrem Einsatz zur Überprüfung von Kompetenzmodellen empirisch erprobt und u. a. nach psychometrischen Kriterien selektiert (Grube, 2011; Schmiemann, 2010). Die Population, an der Kompetenzmodelle überprüft wurden, setzte sich in den bik-Projekten aus Schülerinnen und Schülern der Sekundarstufe I allgemeinbildender Schulen, häufig in

einem querschnittlichen Design, zusammen. Die so erhobenen Daten wurden zur Überprüfung des entsprechenden Kompetenzmodells, häufig mit Verfahren der *Item Response Theory* (IRT), herangezogen (Eggert, 2008; Grube, 2011; Schmiemann, 2010). Als Messmodelle wurden häufig das Rasch-Modell bzw. multidimensionale IRT-Modelle gewählt. Zusammenhänge mit anderen Variablen wurden korrelativ, varianz- oder regressionsanalytisch aufgeklärt (Grube, 2011; Kramer, 2009; Schmiemann, 2010).

In Anlehnung an diese Konzeption von Projekten, die ähnliche Zielstellungen verfolgen, werden im hier vorgestellten Projekt Schülerinnen und Schülern allgemeinbildender Schulen in den Jahrgangsstufen vor dem MSA (Jahrgangsstufe 7 bis 10) einbezogen. Das Testinstrument, das in dieser Zielgruppe eingesetzt werden soll, sollte die Analyse mit IRT-Modellen erlauben. Da hier nicht Denkprozesse im Fokus stehen, sondern eher konzeptuelles Wissen betrachtet wird, bietet sich ein geschlossenes Antwortformat an.

2.2 Modelle

Um Modellkompetenz zu beschreiben, sind wissenschaftstheoretische Konzepte zu Modellen wesentlich. Im Alltag wird der Begriff *Modell* mit unterschiedlichen Bedeutungen verwendet. So gelten sowohl Spielzeugautos als auch Computersimulationen oder Laufstegmodels als Modelle. Objekte, die als Modelle aufgefasst und verschiedenen Bereichen zugeordnet werden, weisen auch in der Wissenschaft keine gemeinsamen Eigenschaften auf, die sie als Modell auszeichnen (Mahr, 2003; Mittelstraß, 2004; Stachowiak, 1973). Oh und Oh (2011, S. 1112) fassen die Debatte um den Modellbegriff zusammen: *„(…) no unique definition of a model is established*".

Der Ursprung des Begriffs „Modell" ist etwa 3000 Jahre v. Chr. das indogermanische **me(d)* mit der Bedeutung „abstecken, messen". Während er vor etwa 2000 Jahren eine konkrete, relative Maßgröße bezeichnete (*modulus*), entwickelte er später eine zunehmend abstrakte Bedeutung und wurde auf zweidimensionale (*modul*, *modellus*) bzw. dreidimensionale Architekturmodelle (*modello*) angewendet (Mahr, 2003). Damit beginnt das Verständnis vom Modell als Träger und Transporteur von Vorstellungen und als Medium von

Ideen, das bis heute gilt (Mahr, 2003, 2008b; Stachowiak, 1973). Ein Laufstegmodel transportiert zum Beispiel Modevorstellungen eines Designers, ein Spielzeugauto eine Vorstellung über Merkmale eines „echten" Autos.

Vereinfacht kann man sagen, dass dieser „Transport" darauf beruht, dass ein Modell etwas anderes repräsentiert (Oh & Oh, 2011). Dieses „repräsentierte Etwas" wird in der Didaktik der Biologie in der Regel als Original bezeichnet und für den schulischen Kontext häufig vereinfacht als Realobjekt betrachtet (Kattmann, 2006). Wissenschaftstheoretische Texte verwenden hierfür häufig den Begriff „*referent*" oder „*target*" (Oh & Oh, 2011) bzw. „Objekt" (Mahr, 2008a) und fokussieren einen konstruktivistisch geprägten Ansatz: Wirklichkeit kann nicht entdeckt, sondern nur konstruiert werden (Stachowiak, 1983; van Fraassen, 1980). Die Klärung des Originalbegriffs ist dementsprechend theoretische Grundlage dieser Arbeit (Kapitel 2.2.1). Bereits bei der Wahrnehmung empirischer Phänomene spielen Modelle eine Rolle: Ein Gegenstand ist „immer ein aufgefasster Gegenstand" (Mahr, 2008b, S. 24), so dass auf der Grundlage der Wahrnehmung eines Phänomens sowie ontologischer und epistemologischer Grundannahmen ein mentales Modell dieses Phänomens gebildet wird (Johnson-Laird, 1983; Nersessian, 1992, 1999; Vosniadou, 2002). Dieses mentale Modell wird als Original bezeichnet, das als wahrgenommenes Phänomen den Ausgangspunkt für die Entwicklung konzeptueller, d. h. gegenständlicher, externer Modelle bildet (Norman, 1983). Diese haben sowohl gegenständliche Eigenschaften, die Mahr (2008a) dem „Modellobjekt" zuschreibt, als auch eine gedankliche Grundlage, die Mahr (2008a) als „Modell" bezeichnet.

Wie ein Original in einem konzeptuellen Modell abgebildet wird, ist eine weitere theoretische Grundlage für das hier vorgestellte Projekt (Kapitel 2.2.2). Indem ein ausgewählter gedanklicher Inhalt vom Original auf das Modell übertragen wird, ist ein konzeptuelles Modell nicht mehr nur ein *Gegenstand für sich*, sondern kann als *Modell von etwas* das Original repräsentieren (Mahr, 2008a, 2008b). Watson und Crick übertrugen z. B. die Struktur der Desoxyribonukleinsäure (DNA) auf ein Modell, das sie repräsentiert. Da der Modellierer darüber entscheidet, welche Merkmale des Originals hierfür relevant sind, ist

die Entwicklung und Nutzung von Modellen zur Repräsentation von Phänomenen immer zweckgebunden und intentional (Giere, 2001, 2009; Mahr, 2008a, 2008b; Stachowiak, 1983). Damit etwas als Modell klassifiziert werden kann, fordert Mahr (2008a, 2008b) darüber hinaus, dass es als *Modell für etwas* zur Erkenntnisgewinnung über das Original eingesetzt werden kann[6].

Entsprechend muss diese Arbeit auch die Perspektive auf Modelle als Methode im wissenschaftlichen Erkenntnisprozess thematisieren (Kapitel 2.2.3). Modelle ermöglichen, die Gültigkeit von Theorien zu beurteilen und somit Erkenntnisse über Originale zu gewinnen, indem auf der Grundlage von Analogien die Passung zwischen dem Modell, der Theorie und der Empirie beurteilt wird (Mahr, 2009; Stachowiak, 1973). Erst durch die doppelte Beziehung zum Original als *Modell von* und *für etwas* wird etwas als Modell aufgefasst (Mahr, 2008a, 2008b). Damit wird die traditionell retrospektive Auffassung eines Modells als Abbild des Originals um eine prospektive Sicht erweitert.

Wie sich diese mediale wie methodische Nutzung von Modellen in der Wissenschaft im schulischen Kontext niederschlägt, ist eine weitere relevante Grundlage für das hier vorgestellte Projekt (Kapitel 2.2.4). Wenn Modelle als Medien eingesetzt werden, wird vor allem *learning science* (Hodson, 1993), d. h. das Lernen von Fachwissen, als Ziel naturwissenschaftlichen Unterrichts erreicht (Gobert & Pallant, 2004; Henze, Van Driel & Verloop, 2007; Schwarz & White, 2005). Das Potenzial von Modellen im schulischen Kontext wird jedoch erst mit der methodischen Nutzung von Modellen als Mittel der Erkenntnisgewinnung ausgeschöpft, da sowohl mit der Entwicklung als auch mit der Anwendung von Modellen spezifische Lernprozesse verbunden sind. Die Reflexion über Modelle und den Prozess der Modellbildung trägt im naturwissenschaftli-

---

[6] Mahr (2008a, 2008b) spricht nicht explizit von einer Klassifizierung als Modell, sondern vom Urteil des ModellseinS. Für dieses Urteil beschreibt er aber relevante Perspektiven, die als Kriterien verstanden werden können.

chen Unterricht dazu bei, die Ziele *learning about science* und *learning how to do science* (Hodson, 1993) zu erfüllen (Henze et al., 2007), d. h. ein Wissenschaftsverständnis sowie manuelle Fertigkeiten zu erlangen.

2.2.1 Mentale Modelle als Zugang zu empirischen Phänomenen

Der Begriff „Phänomen" wird als Sammelbegriff für alle relativ stabilen, allgemeinen Eigenschaften der Empirie verwendet, die aus einer wissenschaftlichen Perspektive Fragen aufwerfen (Frigg & Hartmann, 2012). Modelle vereinfachen als Interpretationen empirischer Phänomene den Zugang zu ihnen, indem sie fokussiert relevante Aspekte von Phänomenen bündeln und so den gedanklichen Aufwand im Umgang mit ihnen reduzieren (Bailer-Jones, 2000; Nersessian, 1992).

Johnson-Laird (1983) argumentiert, dass die Wahrnehmung eines Phänomens durch Sinneseindrücke zur Entwicklung eines Modells dieses Phänomens genutzt wird. Auch Mahr beschreibt, dass ein Gegenstand „immer ein aufgefasster Gegenstand" (Mahr, 2008b, S. 24) ist. Damit hat bereits der Ausgangspunkt der Modellierung als interne, kognitive Repräsentation eines empirischen Phänomens modellhaften Charakter und das Original entspricht dem wahrgenommenen Phänomen (Buckley & Boulter, 2000). Johnson-Laird (1983, S. 406) fasst zusammen: *„Indeed, perception provides us with our richest model of the world*." Dieses Modell versteht er als mentale Repräsentation, die die Struktur eines Phänomens mit räumlichen, zeitlichen und kausalen Beziehungen abbildet (vgl. Nersessian, 1992; Vosgerau, 2006; Vosniadou, 2002). Hierdurch sind mentale Modelle domänenspezifisch (Greca & Moreira, 2000). Wie sie trennscharf von propositionalen Repräsentationen und mentalen Bildern (Johnson-Laird, 1983) abzugrenzen sind und welches Format sie haben (propositional, rein wahrnehmungsbasiert oder nicht-propositional, aber amodal), sind noch offene Fragen (Nersessian, 1999).

Während über die strukturelle Übereinstimmung zwischen mentalem Modell und Phänomen Einigkeit besteht, ist der Charakter dieser Beziehung noch nicht geklärt. Wie mentale Modelle mit empirischen Phänomenen verbunden sind, ist eine offene Frage in der philosophischen Debatte (Held, Knauff &

Vosgerau, 2006; Vosgerau, 2006). Zentral sind hierbei die Fragen, warum es zu fehlerhaften Repräsentationen kommen kann und warum Repräsentationen nicht bidirektional sind, so dass *representandum* und *representans* austauschbar sind. Hierauf gibt es bislang letztlich keine überzeugenden Antworten (Held et al., 2006; Vosgerau, 2006). Trotz aller offenen Fragen sind mentale Modelle aus der Psychologie und Philosophie kaum wegzudenken, da menschliches Denken ohne mentale Modelle kaum erklärt werden kann (Held et al., 2006).

In Abgrenzung zur Theorie der mentalen Logik setzen mentale Modelle zur Erklärung menschlichen Denkens kein Wissen über formale Inferenzregeln voraus: Der Denkende konstruiert und verändert mentale Modelle nicht nach abstrakten logischen Regeln, sondern in Beziehung zur Empirie, die durch mentale Modelle repräsentiert wird (Johnson-Laird, 1983; Nersessian, 1992; Vosgerau, 2006). Damit gründen mentale Modelle ausschließlich in der Wahrnehmung von etwas, ohne dass spezifisches Vorwissen benötigt wird (Vosgerau, 2006).

Bereits entwickelte mentale Modelle können im Langzeitgedächtnis gespeichert werden und beeinflussen die Entwicklung neuer mentaler Modelle (Vosniadou, 2002). Außerdem bilden ontologische und epistemologische Grundannahmen einen Rahmen für die Wahrnehmung. Sie betreffen die Fragen, von welchen Entitäten wir annehmen, dass sie existieren, wie diese kategorisiert werden (ontologische Grundannahmen) und welche Natur unser Wissen hat (epistemologische Grundannahmen; Vosniadou & Ioannides, 1998). Neben dem Phänomen als empirischem Element gehen demnach theoretische Elemente in die Bildung eines mentalen Modells ein (Buckley & Boulter, 2000; Johnson-Laird, 1983; Norman, 1983; Vosniadou & Ioannides, 1998; Vosniadou, 2002; Abb. 2). Dadurch nimmt es eine Mediatorfunktion zwischen beiden ein (Nersessian, 1992, 1999; Vosniadou, 2002).

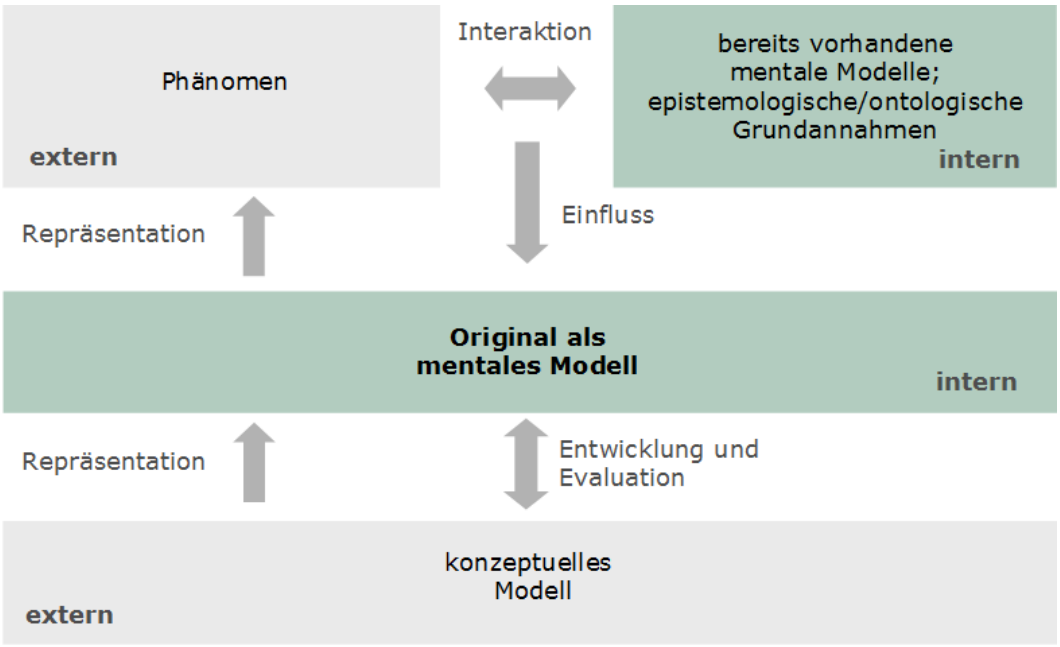


Abb. 2: Mentale Modelle als Zugang zu empirischen Phänomenen und Grundlage für konzeptuelle Modelle.

Ein mentales Modell wird iterativ gebildet, um Anforderungen einer spezifischen Situation, in der es entsteht, zu begegnen (Justi & Gilbert, 2002; Norman, 1983; Vosniadou, 2002). Dies kann z. B. die Lösung eines Problems sein. Mentale Modelle werden vor allem entwickelt, wenn diese Anforderungen nicht durch den Abruf von vorhandenen mentalen Modellen erfüllt werden können. Sie beruhen häufig auf implizitem, verkörpertem Wissen, das in mentalen Modellen explizit wird. Mentale Modelle dienen der Generierung von Erklärungen und Vorhersagen sowie der Interpretation und Akquisition neuer Informationen. Somit mediieren sie die Entwicklung neuer sowie die Revision existierender Theorien (Greca & Moreira, 2000; Nersessian, 1992; Norman, 1983; Vosniadou, 2002). Da neue Information in mentale Modelle integriert werden kann, sind sie immer unvollständig, dynamisch und in hohem Maße adaptiv (Greca & Moreira, 2000; Justi & Gilbert, 2002; Nersessian, 1999). Sie werden außerdem als potenziell inkonsistent betrachtet (Greca & Moreira, 2000; Norman, 1983). Damit deckt sich die Art und Weise, in der mentale Modelle beschrieben werden, mit der Verwendung des Begriffs „Vorstellungen“

von z. B. Gropengießer (2001). Dieser definiert Vorstellungen als „subjektive, gedankliche Konstrukte" (Gropengießer, 2001, S. 31).

Nersessian (1992) unterscheidet drei Formen modellbasierten Denkens – analoges Modellieren, visuelles Modellieren und Gedankenexperimente –, die sie als charakteristisch für das menschliche Denken einordnet (vgl. Vosniadou, 2002). Die Fähigkeit hierzu wird in domänenspezifischen Lernsituationen entwickelt (Nersessian, 1999). In Abgrenzung zu Analogien und Gedankenexperimenten werden beim visuellen Modellieren konzeptuelle Modelle entwickelt, die die Beanspruchung des Gedächtnisses bei der Beantwortung einer Fragestellung oder beim Lösen eines Problems reduzieren, indem sie zusammenhängende Informationen bündeln, von spezifischen Aspekten des Phänomens auf die jeweils relevanten abstrahieren und so die Aufmerksamkeit lenken (Nersessian, 1992, 1999; Norman, 1983). Beim visuellen Modellieren wird zunächst gedanklich ein „Modell" entwickelt, das die Grundlage für ein externes, konzeptuelles „Modellobjekt" bildet (Mahr, 2009). Die Grundlage für ein konzeptuelles Modell bildet somit ein ausgewählter Ausschnitt eines mentalen Modells. Sowohl mentale als auch konzeptuelle Modelle vermitteln zwischen Theorie und Empirie. Sie repräsentieren beide, sind aber funktionell autonom (Morrison & Morgan, 1999a).

Bei der Rezeption konzeptueller Modelle werden wiederum Elemente eines Phänomens, die als relevant eingeschätzt werden, in ein entsprechendes mentales Modell überführt und mit bestehendem Wissen verbunden. Das rezipierte mentale Modell muss nicht notwendigerweise mit dem ursprünglichen mentalen Modell identisch sein, das dem konzeptuellen Modell zugrunde liegt, auch wenn eine Beziehung zwischen dem Phänomen, dem entsprechenden mentalen Modell und dem konzeptuellen Modell besteht (Greca & Moreira, 2000; Justi & Gilbert, 2002; Norman, 1983; Abb. 2). Konzeptuelle Modelle unterstützen jedoch die Entwicklung eines mentalen Modells und den Umgang mit ihm, während sie gleichzeitig auf mentale Modelle aufbauen (Abb. 2). Mentale Modelle können durch konzeptuelle Modelle kommunizierbar gemacht werden, so dass sie von einer *community* geteilt werden können (Nersessian, 1992, 1999). Konzeptuelle Modelle werden deshalb als präzise, in Bezug auf einen

gesetzten Fokus vollständige Repräsentationen verstanden, die mit dem wissenschaftlich akzeptierten Wissen übereinstimmen (Greca & Moreira, 2000; Norman, 1983).

2.2.2 Modell als Medium – Intentionale Abbildung eines Originals

Eine weitere Form modellbasierten Denkens, und zwar Analogien, spielt bei der Entwicklung neuer Gedanken eine zentrale Rolle: Analogien geben neuen Begriffen eine Bedeutung, indem sie eine Brücke zwischen einem existierenden und einem neuen konzeptuellen Rahmen schlagen (Nersessian, 1992). Ausgehend von bekannten Inhalten, der Ausgangsdomäne, werden Verbindungen zu unbekannten Inhalten, der Zieldomäne, gezogen. Dieser Vergleich von Ausgangs- und Zieldomäne betrifft nur Teile der Strukturen der beiden Domänen (Duit, 1991; Gentner, 1989, 2002). Eine Analogie ist demnach ein Weg, auf Gemeinsamkeiten voneinander unabhängiger Systeme zu fokussieren („*structure-mapping*", Gentner, 1989). Durch die Abstraktion von der Ausgangsdomäne werden modellhaft Inferenzschlüsse auf die Zieldomäne gezogen, so dass die Anforderungen der jeweiligen (Problemlöse-)Situation erfüllt werden (Nersessian, 1992).

Analogien zwischen Elementen eines Modells und eines Originals parallelisieren Modell und Original, so dass ein konzeptuelles Modell ein Original repräsentiert (Giere, 2006, 2009; Stachowiak, 1973). Die meisten Autoren nehmen an, dass es nur partielle Übereinstimmungen zwischen Modell und Original gibt, da nicht alle Eigenschaften des Originals im Modell repräsentiert werden (z. B. da Costa & French, 2000; Giere, 1999, 2006, 2009). Zum Charakter dieser Übereinstimmungen gibt es verschiedene Positionen: Sie werden entweder als Isomorphien, d. h. umkehrbar eindeutige Abbildungen von Originaleigenschaften in Modelleigenschaften (da Costa & French, 2000; van Fraassen, 1980), oder als Ähnlichkeitsbeziehungen (Giere, 1999, 2006, 2009) verstanden. Hier besteht weiterer Forschungsbedarf, um zu klären, wodurch Modelle etwas repräsentieren, auf welche Weise sie etwas repräsentieren und welche Kriterien für die Adäquatheit einer wissenschaftlichen Repräsentation gelten (Frigg, 2006; Frigg & Hartmann, 2012).

Einigkeit besteht hingegen darin, dass die Repräsentation eines Originals durch ein konzeptuelles Modell intentionale Anteile hat. Da die Analogien zwischen Modell und Original von einem Subjekt selektiv gezogen werden, beschreibt Stachowiak (1973, 1983) die Repräsentation eines Originals durch ein Modell als subjektiv und perspektivgebunden: Indem die Erschließung der Empirie durch Modelle „auf das – passive oder aktive – Erfassen von etwas aus ist, vollzieht sie sich relativ zu bestimmten Subjekten, ferner selektiv – intentional selektierend und zentrierend – und in je zeitlicher Begrenzung ihres Original-Bezuges" (Stachowiak 1983, S. 56). Giere (2001, 2009) bezeichnet die zweckgebundene Entwicklung und Nutzung von Modellen zur Repräsentation von Phänomenen als „*intentional aspect of symbolic communication*" (Giere, 2009, S. 275).

Konzeptuelle Modelle sind demnach nicht nur Modelle von etwas (einem Original), sondern auch Modelle für jemanden. Stachowiak (1973) bezeichnet dies als pragmatischen Entschluss des Modellierers darüber, welche Merkmale des Originals relevant sind und somit erfasst bzw. hervorgehoben (Kontrastierung, Stachowiak, 1973, 1983) werden. Entsprechend weist ein Teil der Eigenschaften des Modells keine Ähnlichkeitsrelation zu denen des Originals auf, da es nicht alle Originaleigenschaften beinhaltet (präterierte Attribute, Stachowiak 1973, 1983). Darüber hinaus hat es als Modellobjekt neue Eigenschaften, die das Original nicht hat (Mahr, 2008a; abundante Attribute, Stachwowiak, 1973, 1983). Der Modellierer nutzt das Modell demnach nicht dazu, das Original zu spiegeln, sondern von ihm zu abstrahieren oder es zu übersetzen (Morrison & Morgan, 1999a). Er beantwortet somit die Frage "Wovon, für wen, wann und wozu ist etwas ein Modell?" (Stachowiak, 1973). In Passung hierzu formuliert Stachowiak (1973, S. 131) drei Hauptmerkmale von Modellen:

1. *Abbildungsmerkmal:* Modelle sind immer Modelle von etwas, das heißt, sie repräsentieren Originale.
2. *Verkürzungsmerkmal:* Modelle erfassen nur die Merkmale des Originals, die für den Modellierer bzw. Verwender des Modells relevant sind.

3. *Pragmatisches Merkmal:* Modelle sind Originalen nicht eindeutig zugeordnet, sondern erfüllen eine Ersetzungsfunktion für bestimmte Subjekte, in bestimmten Zeitintervallen und zu einem bestimmten Zweck.

Auch wenn Stachowiak bereits herausarbeitet, dass Modelle von einem Subjekt intentional geprägt sind (vgl. Giere, 2001), beziehen sich die Merkmale, mit denen er Modelle beschreibt, ausschließlich auf die Herstellung eines Modells von etwas. Zudem suggerieren sie, dass anhand von Objekteigenschaften entschieden werden kann, ob etwas ein Modell ist – dies ist jedoch aufgrund der vielfältigen Verwendung des Modellbegriffs (s. o.) nicht möglich.

Mahr (2009) formuliert stattdessen einen dynamischen, funktionalen Modellbegriff. Ein Modell kann zunächst als Modellobjekt als ein „*Gegenstand für sich*" (Mahr, 2004, S. 11, Hervorhebung im Original) angesehen werden, der eine beliebige Erscheinungsform aufweist. Darüber hinaus erfüllen Modelle nach Mahr (2004, 2008a, 2008b) mit Blick auf das Original zwei Funktionen. Sie können zum einen als *Modelle von etwas* betrachtet werden, wenn Wissen über das Original, z. B. aus Beobachtungen oder Experimenten, induktiv in ein Modell eingebracht wird. Sie sind somit Ergebnis einer Induktion vom Original aus. Hierzu analysiert der Modellierer, welche Teile von Modell, Theorie und Empirie wie zusammenpassen. Dabei lernt er bereits etwas über das zugrundeliegende Phänomen und bildet dazu eine Theorie (Morrison & Morgan, 1999a). Modelle können darüber hinaus auch als *Modelle für etwas* betrachtet werden, wenn durch die Anwendung des Modells Erkenntnisse über das Original deduktiv abgeleitet werden können. Damit ein Objekt als Modell aufgefasst werden kann, muss es demnach nicht nur ein Original intentional abbilden, sondern auch zur Gewinnung von Erkenntnissen über das Original angewendet werden können (Mahr, 2004, 2008a, 2008b).

2.2.3 Modell als Methode – Entwicklung von Theorien und Gewinnung von Erkenntnissen

Die Grundlage für die elaborierte, intentionale Nutzung von Modellen zur Erkenntnisgewinnung in der Wissenschaft wird durch die Nutzung mentaler Modelle beim Wissenserwerb während der individuellen kognitiven Entwicklung

gelegt (Vosniadou, 2002). Dort werden Modelle nach dem *model-based account of scientific theories* (auch *semantic view*, s. u.) als Basis für die wissenschaftliche Theoriebildung gesehen (Cartwright, 1983; Giere, 2001). Hiermit sind wissenschaftstheoretische Fragen verbunden, die sich darauf beziehen, in welcher Beziehung Modelle zu Theorien stehen und auf welche Weise mit Modellen Erkenntnisse gewonnen werden.

Die wissenschaftstheoretische Betrachtung von Modellen hat damit zwei Schwerpunkte (Bailer-Jones, 1999): zum einen die Abgrenzung von Modell und Theorie mit Blick auf die Entwicklung von Theorien in einem formalen, modelltheoretischen Ansatz (*semantic view of scientific theories*; z. B. Giere, 1988; van Fraassen, 1980), der auf den logischen Positivismus aufbaut (*received* oder *syntactic view of theories*; z. B. Carnap, 1928), zum anderen die Betrachtung der Funktion von Modellen im wissenschaftlichen Erkenntnisprozess mit Blick auf die Änderung von Theorien in einem funktionalen Ansatz (z. B. Hesse, 1966).

Formaler Ansatz: Modelle und Theorien

In einer wissenschaftstheoretischen Position der 1920er bis 1930er Jahre, dem logischen Positivismus, wurden Modelle zunächst im *syntactic view of theories* als Anhängsel von Theorien mit höchstens pädagogischem oder ästhetischem Wert verstanden. Nach dieser Position werden Theorien formal in einer Sprache aus empirischen und theoretischen Termen rekonstruiert. Die empirischen Terme beschreiben beobachtbare Phänomene, während theoretische Terme ihre Bedeutung durch ihre beobachtbaren Implikationen erhalten (Carnap, 1928). Ein Problem dieser Position ist, dass theoretischen Termen in der Regel mehr als eine Bedeutung zugeschrieben werden kann (Morrison & Morgan, 1999b).

Die Vertreter des *semantic view of scientific theories* (z. B. Giere, 1988; van Fraassen, 1980) lehnen eine formale Analyse von Theorien, wie sie im logischen Positivismus vorgenommen wurde, ab. Stattdessen rücken sie den prozesshaften Charakter von Wissenschaft stärker in den Vordergrund und sehen

Modelle zunehmend als wichtigen Bestandteil hiervon. Sie betrachten Modelle als zentrale Einheit wissenschaftlicher Theoriebildung und verstehen Modelle als nicht-sprachliche mengentheoretische *Strukturen* (Strukturalismus; van Fraassen, 1980). Der Zusammenhang zwischen Theorie und Modell wird im *semantic view* unterschiedlich konzeptualisiert: Der *instantial view* versteht Modelle als konkrete Instantiierungen von abstrakt beschriebenen Axiomen von Theorien, d. h. Modelle interpretieren Theorien (Suppes, 1961). Der *representational view* begreift Modelle als Repräsentationen nicht-linguistischer Strukturen, die die Welt abbilden (Giere, 1999). Der *intentional view* betont, dass Modelle in Abgrenzung zu Theorien die intendierte Anwendung als integralen Bestandteil beinhalten (Suárez, 1999).

Neben dieser Betrachtung des Zusammenhangs von Theorie und Modell unter unterschiedlichen inhaltlichen Foci können sie allgemein im Abstraktionsgrad unterschieden werden. Dieser Unterschied kann 1. als gradueller Unterschied zwischen unterschiedlich abstrakten Modellen (Giere, 1999, 2006), 2. als hierarchische Beziehung mit Modellen als Bausteinen von Theorien (Develaki, 2007; Hesse, 1966) oder 3. als funktionelle Autonomie von Modellen gegenüber Theorien (Morrison & Morgan, 1999a, 1999b) verstanden werden. Nach dem Verständnis als *graduellem Unterschied* formuliert die Wissenschaft ausgehend von sehr abstrakten, übergreifenden Modellen wie z. B. der Evolutionstheorie spezifischere Modelle, die Aspekte der Welt repräsentieren. Theorien haben diesem Verständnis nach die Funktion, Zusammenhänge zwischen Elementen der spezifischeren Modelle zu charakterisieren, und müssen erst durch spezifischere Modelle auf die Welt bezogen werden (Giere, 1999, 2006). Wenn Modelle als *Bausteine von Theorien* betrachtet werden, werden Theorien als Gruppen von Modellen verstanden. Theorien repräsentieren dabei über den empirischen Anteil der jeweiligen Modelle Phänomene (van Fraassen, 1980). Da theoretische wie empirische Elemente in der Entwicklung von Modellen herangezogen werden und man mit Modellen sowohl etwas über Theorien als auch über die Welt lernen kann, schreiben Morrison und Morgan (1999a) ihnen eine partielle Unabhängigkeit und *funktionelle Autonomie* gegenüber Theorie und Empirie zu. Damit verbinden sie den Gedanken, dass Modelle zwischen Theorien und der Welt mediieren. Der Unterschied zwischen Modellen

und Theorien besteht nach dieser Konzeption darin, dass Theorien sich aus allgemeinen Prinzipien zusammensetzen, die das Verhalten großer Gruppen von Phänomenen beschreiben, während Modelle inhaltlich begrenzter sind. In der Regel werden mehrere Modelle benötigt, um diese Prinzipien anzuwenden.

Die verschiedenen Positionen zum Verhältnis von Modellen und Theorien stimmen darin überein, dass sich Modelle und Theorien dadurch unterscheiden, wie stark sie von Phänomenen abstrahieren. Während Theorien ein übergreifender Charakter zugeschrieben wird, haben Modelle die Funktion, einzelne Phänomene bzw. Phänomenausschnitte spezifisch zu repräsentieren. Damit spezifizieren Modelle über die Theorie hinausgehende Bedingungen, wie eine Theorie konkret auf die empirische Welt angewendet werden kann. Anders als im *syntactic view* (z. B. Carnap, 1928), der Phänomene und Theorien direkt aufeinander bezieht, vermitteln Modelle nach dem *semantic view* (z. B. Giere, 1988; van Fraassen, 1980) zwischen beiden.

Funktionaler Ansatz: Modelle im Prozess der Erkenntnisgewinnung

Mit einer wissenschaftstheoretischen Position ist auch eine erkenntnistheoretische verbunden. Diese kann sowohl realistisch sein, d. h. wie etwa der wissenschaftliche Realismus davon ausgehen, dass Theorien *wahre* Aussagen über die Realität treffen können (Devitt, 2005), als auch nicht-realistisch wie z. B. der konstruktive Empirismus. Dieser nimmt Theorien als gültig an, wenn sie *empirisch adäquat* sind, d. h. korrekt beschreiben, was beobachtbar ist, und alle Phänomeneigenschaften mit Teilen ihrer Modelle identifizierbar sind (van Fraassen, 1980).

Unabhängig davon, ob man Theorien als wahre oder empirisch adäquate Aussagen auffasst, spielen Modelle eine Rolle bei der Beurteilung der Gültigkeit von Theorien. Der hypothetische Charakter von Modellen bildet die methodische Grundlage für die Formulierung von Theorien mit allgemein verbindlichem Anspruch und ermöglicht die Gewinnung von Erkenntnissen über die Welt (Abel, 2008; Stachowiak, 1973): Ein ausgewählter gedanklicher Inhalt wird vom Original auf das Modell übertragen, verändert sich bei der zielgerich-

teten Anwendung des Modells und führt beim hypothetischen Rückschluss auf das Original zu dessen Veränderung – die Vorstellung des Phänomens verändert sich durch neue Informationen, die mit dem Modell gewonnen wurden (Stachowiak, 1973; Mahr, 2009). Dabei wird aufbauend auf die Analogie zwischen Elementen des Modells und Elementen der Welt eine Hypothese über die Passung zwischen dem Modell und der Welt getestet (Giere, 2001, 2004). Da man häufig nur indirekt mit wissenschaftlichen Untersuchungsgegenständen interagieren kann und man somit nicht unmittelbar sieht, ob ein Modell passt, ist dies die einzige Möglichkeit, mit Modellen Erkenntnisse über die Welt zu gewinnen (Giere, Bickle & Mauldin, 2006). Das Modell vermittelt auf diese Weise zwischen Theorie und Empirie und ist nicht nur Medium für Inhalte des Originals, sondern auch Methode zur Erkenntnisgewinnung (Mahr, 2009; Morrison & Morgan, 1999a; Stachowiak, 1973; Abb. 3). Ein konzeptuelles Modell wird in unterschiedlichen Wissenschaftsbereichen, etwa den Natur- und Gesellschaftswissenschaften, unterschiedlich getestet. Das Vorgehen hierzu ist somit domänenspezifisch (Oh & Oh, 2011).

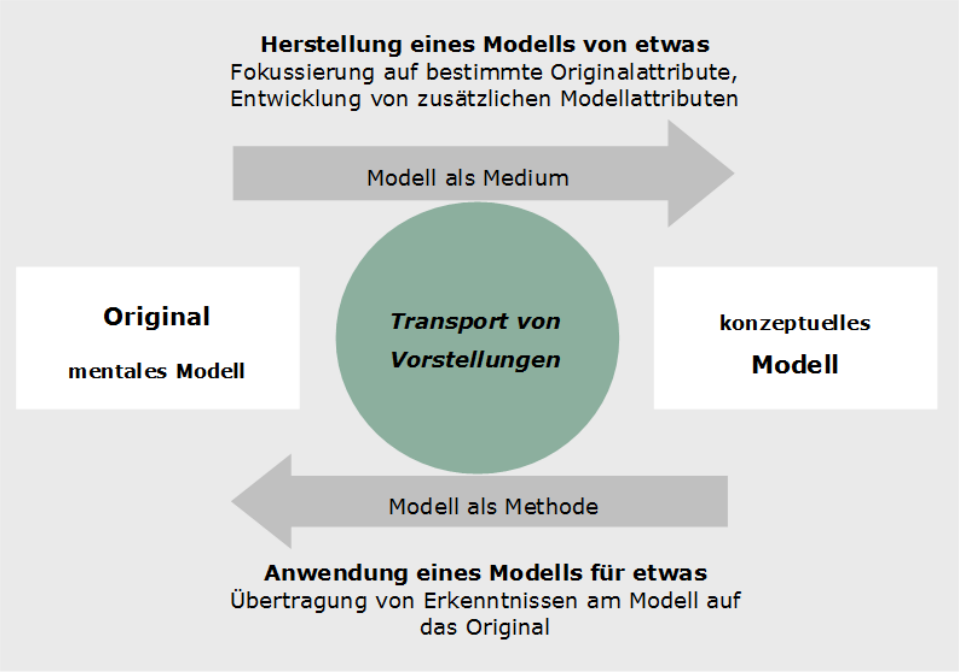


Abb. 3: Beziehungen zwischen Original und konzeptuellem Modell.

2.2.4 Mediale und methodische Nutzung von Modellen in der Schule

In der Wissenschaft spielen Modelle eine wichtige Rolle – Gilbert (1991) definiert Wissenschaft sogar als Konstruktion von Modellen, welche Konzepte der Welt repräsentieren. Da Modelle sowohl Produkt als auch Methode der Wissenschaft sind, ist diese ohne Modelle weder lehr- noch lernbar (Harrison & Treagust, 2000). Sie sind folglich Bestandteil des Schulunterrichts, da Kompetenz im Umgang mit Modellen sowohl im Berufsleben als auch im Sinne einer *scientific literacy* relevant ist (Kapitel 1). Ein Ziel naturwissenschaftlichen Unterrichts ist mit *learning science* (Hodson, 1993) der Erwerb von Fachwissen (Oh & Oh, 2011). Es kann erreicht werden, wenn Modelle als Medien eingesetzt werden (Gobert & Pallant, 2004; Henze et al., 2007; Schwarz & White, 2005). Lehrerinnen und Lehrer können mit Modellen demonstrieren, wie etwas funktioniert, und komplexe wissenschaftliche Konzepte anschaulich erklären, so dass sie für Schülerinnen und Schüler zugänglich werden (Oh & Oh, 2011). Dadurch wird die Bildung entsprechender mentaler Modelle unterstützt (Buckley & Boulter, 2000; Nersessian, 1999; Oh & Oh, 2011). Der Umgang mit alternativen Modellen ermöglicht die Repräsentation verschiedener Aspekte eines Phänomens, so dass unterschiedliche Bedürfnisse und Voraussetzungen der Schülerinnen und Schüler berücksichtigt werden können. Darüber hinaus schränkt die Repräsentation des gleichen Inhalts durch verschiedene Modelle die Wahrscheinlichkeit ein, dass Schülerinnen und Schüler mentale Modelle zu einem davon abweichenden Inhalt entwickeln (vgl. Kapitel 2.2.1), und führt durch Abstraktionsprozesse zu einem tieferen Verständnis des Inhalts (Ainsworth, 2008; Buckley & Boulter, 2000). Vosniadou und Ioannides (1998) betonen jedoch, dass die Entwicklung von zunehmend wissenschaftlichen Vorstellungen dadurch begleitet werden sollte, dass Schülerinnen und Schüler sich ihrer vorhandenen Vorstellungen und Grundannahmen bewusst werden und größere theoretische Konstruktionen mit einer größeren Erklärungskraft bilden. Hierfür müssen Modelle auch zum Erwerb von Fachwissen reflektiert werden, um beim Lernen Tiefenverarbeitungsstrategien statt Oberflächenstrategien zu aktivieren (Sins et al., 2009).

Indem man Modelle als Medien einsetzt, wird ihr Potenzial jedoch nicht ausgeschöpft: "*We do not learn much from looking at a model – we learn more*

from building the model and from manipulating it" (Morrison & Morgan, 1999a, S. 12). Sowohl mit der Entwicklung als auch mit der Anwendung von Modellen sind spezifische Lernprozesse verbunden. Diese tragen im naturwissenschaftlichen Unterricht zu zwei weiteren Zielen bei (Henze et al., 2007): mit *learning about science* zur Entwicklung von Wissenschaftsverständnis (Schwarz & White, 2005) und mit *learning how to do science* (Hodson, 1993) zur Entwicklung wissenschaftlichen Denkens (Lehrer & Schauble, 2006). Ein Ansatzpunkt hierzu kann etwa die Thematisierung des Charakters von Modellen als partielle Repräsentationen von Phänomenen sein, bei denen nur einzelne Elemente des Originals auf das Modell übertragen werden (da Costa & French, 2000; Giere, 1999, 2006, 2009). Indem zum einen reflektiert wird, inwiefern die Wahl der Analogien das Modell und seine Erklärungskraft beeinflusst, und zum anderen die Relevanz der Art der Repräsentation diskutiert wird, können Schülerinnen und Schüler ihre Vorstellungen zu Modellen entwickeln. Die Testung ihrer eigenen Ideen mithilfe alternativer Modelle, die zur Veränderung dieser Modelle führen kann, hilft ihnen, sowohl die spezifischen Modelle als auch das wissenschaftliche Vorgehen zu verstehen (Justi & Gilbert, 2006; Schwarz & White, 2005; Schwarz & Gwekwerere, 2007).

2.3 Vorstellungen zu Modellen

Kognitive Strukturen, die das Verständnis von Modellen und dem Umgang mit ihnen abbilden, werden als mentale Modelle bzw. Vorstellungen bezeichnet. Sie werden entwickelt oder aktiviert, um Anforderungen einer Situation zu begegnen (vgl. Kapitel 2.2.1) und gehören zur kognitiven Facette von Kompetenz. Für die Erfassung von Modellkompetenz ist relevant, welche Ansätze zur Erfassung von Vorstellungen zu Modellen in der Forschung bereits genutzt werden (Kapitel 2.3.1). Von kontextunabhängigen Interviews bis zu kontextualisierten Items in geschlossenem Antwortformat kommen hierzu verschiedene Instrumente zum Einsatz (AAAS, o. J.; Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; Schwarz & White, 2005; Treagust, Chittleborough & Mamiala, 2001, 2002, 2004; Trier & Upmeier zu Belzen, 2009; Van Driel & Verloop, 1999).

Grundlage für ein Kompetenzmodell der Modellkompetenz sind neben wissenschaftstheoretischen Positionen (vgl. Kapitel 2.2.2, 2.2.3) empirische Befunde zu Vorstellungen, die Schülerinnen und Schüler zu Modellen haben, da diese sich individuell unterscheiden. Für die Strukturierung und Graduierung von Modellkompetenz ist relevant, wie diese Vorstellungen in der Forschung bislang inhaltlich kategorisiert (Kapitel 2.3.2) und graduiert werden (Kapitel 2.3.3). Kategorien lassen sich in die folgende Bereiche einordnen: Entwicklung von Modellen, Beziehung zwischen Modell und Original, alternative Modelle, Zweck von Modellen, Testen von Modellen sowie Ändern von Modellen. Dabei beziehen sich prominente Vorstellungen auf einen medialen Modellbegriff (AAAS, o. J.; Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; Schwarz & White, 2005; Treagust et al., 2001, 2002; Trier & Upmeier zu Belzen, 2009; Van Driel & Verloop, 1999; vgl. Kapitel 2.2.1). Die Vorstellungen werden in drei bis vier Niveaus graduiert, die von einer medialen bis zu einer methodischen Perspektive auf Modelle reichen. Hierzu beziehen sich Crawford und Cullin (2005), Grosslight et al. (1991), Schwarz et al. (2009) sowie Stephens, McRobbie und Lucas (1999) auf die Komplexität des Wissenschaftsverständnisses, die sich in den Vorstellungen zu Modellen niederschlägt.

2.3.1 Erfassung

Vorstellungen zu Modellen werden mit sehr unterschiedlichen Instrumenten erhoben. Crawford und Cullin (2005), Grosslight et al. (1991) Justi und Gilbert (2003) sowie Trier und Upmeier zu Belzen (2009) nutzten einen halbstrukturierten Interviewleitfaden nach Grosslight et al. (1991). Anhand des Leitfadens führten sie ohne Bezug zu spezifischen Kontexten, d. h. fachlichen Inhalten oder Situationen, Interviews mit Schülerinnen und Schülern (Grosslight et al., 1991; Trier & Upmeier zu Belzen, 2009), Lehrerinnen und Lehrern bzw. Lehramtstudierenden (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003) sowie Expertinnen und Experten (Grosslight et al., 1991; Justi & Gilbert, 2003) durch. Trier und Upmeier zu Belzen (2009) legten den Schülerinnen und Schülern zu Beginn der Interviews Objekte vor, die die Schülerinnen und Schüler daraufhin beurteilen sollten, ob sie Modelle sind. Auf diese Beispiele nahmen die Schülerinnen und Schüler in den Interviews z. T. Bezug.

Die Vorstellungen wurden jedoch unabhängig von spezifischen Situationen erhoben. Die Aussagen wurden in den genannten Interviewstudien nach einem induktiv entwickelten Codierschema kategorisiert, wobei Crawford und Cullin (2005) das von Grosslight et al. (1991) publizierte nutzten.

Crawford und Cullin (2005) setzten zusätzlich zu Interviews einen Fragebogen mit Items in offenem Antwortformat ein, die Interviewfragen nach Grosslight et al. (1991) entsprechen, um schnell einen Überblick über die Vorstellungen der Lehrerinnen und Lehrer zu erhalten. Mit diesen führten sie vor einer Intervention Interviews ohne Kontextbezug und griffen in anschließenden Interviews auf Kontexte der Intervention zurück. Sins et al. (2009) setzten zur Befragung von Schülerinnen und Schülern ebenfalls Items in offenem Antwortformat ein, im Gegensatz zu Crawford und Cullin (2005) jedoch nach einer Intervention und mit einer induktiven Kategorisierung. Eine Gemeinsamkeit der beiden Studien ist die Kontextgebundenheit des post-Tests. Van Driel und Verloop (1999) verwendeten ähnlich wie Crawford und Cullin (2005) in Anlehnung an Grosslight et al. (1991) kontextunabhängige Items in offenem Antwortformat und kategorisierten die Antworten von Lehrerinnen und Lehrern induktiv. Darüber hinaus setzten sie ebenfalls kontextunabhängige Items mit einer vierstufigen Likertskala ein, auf der Lehrerinnen und Lehrer einschätzen sollten, wie zutreffend die präsentierten Aussagen für Modelle und Modellbildung in der Wissenschaft sind.

Auch Treagust et al. (2001, 2002, 2004) arbeiteten quantitative Ansätze mit Likert-skalierten Fragebogenitems aus. Aufbauend auf den VOSTS-Items (*Views of Science-Technology-Society*) von Aikenhead und Ryan (1992) beinhaltet der VOMS (*My Views of Models in Science*) *Forced-Choice* Items (FC-Items), in denen Schülerinnen und Schüler sich zwischen zwei kontextunabhängigen Aussagen über Modelle entscheiden müssen. Ein weiterer Ansatz, der SUMS (*Students' Understanding of Models in Science*; Treagust et al., 2002), basiert auf der Studie von Grosslight et al. (1991) und beinhaltet kontextunabhängig formulierte Items mit fünfstufiger Likert-Skala, mit denen ebenfalls Schülerinnen und Schüler befragt wurden. Als post-Test für eine Interventionsstudie entwickelten Treagust et al. (2004) zusätzlich zum VOMS

den Fragebogen *Molecular Representations*, in dem Schülerinnen und Schüler zu vier Modellen jeweils fünf Aussagen zum Zweck des jeweiligen Modells auf einer fünfstufigen Likert-Skala bewerteten.

Einen weiteren Ansatz, Vorstellungen zu Modellen quantitativ zu erfassen, hat die AAAS im Rahmen des Projekts 2061 mit MC-Items zu Modellen für die Jahrgangsstufen 6 bis 12 entwickelt (AAAS, o. J.). Die Items sind so konzipiert, dass die Antwortmöglichkeiten aus einer wissenschaftlichen Vorstellung, einer nicht-wissenschaftlichen Vorstellung sowie aus zwei Aussagen mit einer Kombination aus beiden bestehen. Die Mehrheit dieser MC-Items ist kontextunabhängig formuliert.

Auch Schwarz und White (2005) nutzen MC-Items, die sie mit der Aufforderung zu einer kurzen Begründung ergänzen. Darüber hinaus setzen sie eine Kategorisierungsaufgabe sowie Richtig/Falsch-Aufgaben mit Begründung ein. Die kontextunabhängigen Aufgaben wurden als prä- sowie als post-Test eingesetzt und mit halbstrukturierten Interviews validiert, die auf Kontexte einer Intervention Bezug nahmen.

Zusammenfassend ist festzustellen, dass Vorstellungen bzw. Wissen zu Modellen mit sehr unterschiedlichen Instrumenten erfasst wurden – von Interviews über Items in offenem Antwortformat, Likert-skalierten Items, FC-Items, MC-Items bis zu Kategorisierungs- und Richtig/Falsch-Aufgaben. Bezüge zu Kontexten wurden vor allem in post-Tests sowie in einzelnen AAAS-Items hergestellt. Einige Autoren stellen in Frage, inwiefern die kontextunabhängige Erfassung von Vorstellungen zu Modellen valide ist, da Schülerinnen und Schüler epistemologische Vorstellungen möglicherweise angebunden an einen Kontext, aber nicht als abstrakten Standpunkt formulieren können oder aber nach reinem Hörensagen komplexe Aussagen in Fragebögen auswählen (Sins et al., 2009). Für die Erhebung kontextgebundener und -unabhängiger Vorstellungen zum Wissenschaftsverständnis berichten Urhahne, Kremer und Mayer (2011), dass diese miteinander in Verbindung stehen und die Vertrautheit mit vielen Kontexten mit elaborierten kontextunabhängigen Vorstellungen einhergeht. Diese Argumente und Befunde sprechen dafür, nach der Unterscheidung von

Leisner-Bodenthin (2006) bzw. Leisner und Mikelskis (2004) neben domänenübergreifendem auch -spezifisches Wissen über Modelle zu erheben (Kapitel 2.4.1). Darüber hinaus verlangt die Kontextspezifität von Kompetenzen (vgl. Kapitel 2.1.1) die domänenspezifische Erhebung von Vorstellungen zu Modellen. Entsprechend muss zur Erhebung von Modellkompetenz ein domänenspezifisches, kontextgebundenes Instrument entwickelt werden. Da Vorstellungen zu Modellen nicht als Denkprozesse, sondern als konzeptuelles Wissen einzuordnen sind, erlauben sie die Entwicklung von Items in einem geschlossenen Antwortformat für die empirische Überprüfung eines entsprechenden Kompetenzmodells. Dies ermöglicht die Befragung einer relativ großen Stichprobe und dadurch die Auswertung mit Modellen der IRT (vgl. Kapitel 2.1.2). Während im hier vorgestellten Projekt MC-Items entwickelt wurden (Kapitel 3), beschreiben parallel Krell und Krüger (2010) sowie Grünkorn und Krüger (2012) Modellkompetenz mit FC-Items bzw. Items in offenem Antwortformat.

2.3.2 Inhaltliche Kategorien

Mithilfe der in Kapitel 2.3.1 vorgestellten Instrumente kategorisieren die Autoren der entsprechenden empirischen Studien Vorstellungen zu Modellen. Diese Kategorisierung ist unterschiedlich, aber in einigen Punkten vergleichbar (Anhang 1). Da die Kategorien sowohl in Studien mit Schülerinnen und Schülern als auch mit Lehrerinnen und Lehrern bzw. Expertinnen und Experten gebildet wurden und der Fokus hier auf der Strukturierung in solche Kategorien liegt, werden diese zu beiden Gruppen vorgestellt. Sie fallen in die Bereiche Entwicklung von Modellen, Beziehung zwischen Modell und Original, alternative Modelle, Zweck von Modellen, Testen von Modellen sowie Veränderbarkeit von Modellen.

Entwicklung von Modellen

Vorstellungen zur Entwicklung von Modellen charakterisieren den Modellbildungsprozess als Umsetzung der Gedanken des Modellierers, als Angleichung des Modellverhaltens an das des Originals oder als iterativen Modellentwicklungsprozess (*designing and creating*, Crawford & Cullin 2005). Die Rolle des

Modellierers kann als passiv, aktiv oder gezielt entscheidend charakterisiert werden (*designing and creating – modeler's role*, Grosslight et al., 1991). Darüber hinaus müssen verschiedene Punkte bei der Modellentwicklung beachtet werden (*designing and creating*, Grosslight et al., 1991):

- Eigenschaften des Originals (*what the real thing is*) – worauf muss man achten (Proportionen, Größe, Form, Zusammenhänge),
- Eigenschaften des Modells (*the model itself*; *should be exact, smaller, proportional*; *depends on the view*; *depends on what is major/minor, basiscs/details, little things/big things, main things; depends on what you're talking about, working on*) – wie soll das Modell aussehen (exakt, kleiner, proportional, abhängig von Blickwinkel, grundlegende Form, wichtige und unwichtige Eigenschaften, Hauptidee, Inhalt, Passung zu Zweck),
- Kriterien für die Modellbildung (*general criteria in model building*) – Einfachheit, Validität und Passung des Modells mit dem Zweck,
- Kommunizierbarkeit des Modellinhalts durch das Modell (*communication*) – was will man mitteilen, ist es klar verständlich und nicht irreführend.

Die Eigenschaften des Modells werden demnach abhängig von der intendierten Nutzung gestaltet. Dies wurde von den Expertinnen und Experten in den Vordergrund gestellt. Die Schülerinnen und Schüler betonten dagegen, dass das Modell dem Original in Größe, Form und Proportionen möglichst ähnlich und für einen selbst und/oder andere verständlich sein sollte (Grosslight et al., 1991).

Beziehung zwischen Modell und Original

Einige Vorstellungen betreffen die Beziehung zwischen Modell und Original und daraus resultierende ‚Eigenschaften von Modellen'. Als Grundlage für Modelle wurden unterschiedlich abstrakter Originale, von einem Objekt bis zu einer abstrakten Idee, genannt (*kinds of models – type of the thing modeled*, Grosslight et al., 1991; *entities*, Justi & Gilbert, 2003) und diesen unterschiedlich nah sein. Dabei können Schülerinnen und Schüler die aktive Rolle des

Modellierers bewusst wahrnehmen (*kinds of models – awareness of modeler*, Grosslight et al., 1991). Sie können ein Modell als exakte Reproduktion, als ähnliche, maßstabsgetreue Repräsentation des ganzen bzw. von Teilen des Originals oder als mentales Bild vom Original sehen (*kinds of models – relationship*, Grosslight et al., 1991; *nature*, Justi & Gilbert, 2003). Als Beispiele für Modelle nannten sie Objekte, Personen, visuelle Modelle auf Papier oder dem Computer, verbale Modelle und abstrakte Modelle im Sinne einer Idee oder eines theoretischen Modells (*kinds of models – examples of models*, Grosslight et al., 1991). Van Driel und Verloop (1999) stellten fest, dass vorgegebene Gegenstände sehr unterschiedlich als Modell oder Nicht-Modell klassifiziert werden.

Vorstellungen, die in diese Kategorien zur Beziehung zwischen Modell und Original eingeordnet wurden, thematisierten die Schülerinnen und Schüler sowie die Lehrerinnen und Lehrer stark (Justi & Gilbert, 2003; Van Driel & Verloop, 1999). Dabei formulierten sie z. T. widersprüchliche Vorstellungen (Justi & Gilbert, 2003). Auch die Ergebnisse empirischer Studien liefern ein heterogenes Bild: Grosslight et al. (1991) sowie Schwarz und White (2005) fanden bei den Schülerinnen und Schülern vor allem Vorstellungen vom Modell als Kopie des Originals. Auch die Bearbeitung der AAAS-Items lässt bei der Mehrheit der Schülerinnen und Schüler auf ein solches Modellverständnis schließen, das einen zweckgebundenen Fokus des Modells vernachlässigt (AAAS, o. J.). Dagegen suggerieren Ergebnisse von Treagust et al. (2001, 2002), dass Schülerinnen und Schüler Modelle mehrheitlich für maßstabsgetreue, sehr ähnliche Repräsentationen halten und nur wenige von ihnen Modelle als Kopien sehen.

Nach dem Modellverständnis der meisten Schülerinnen und Schüler sind sowohl Modell als auch Original konkrete Objekte, die sich in Farbe, Form, Größe, Material und Fokussierung vom Original unterscheiden. Die Lehrerinnen und Lehrer sowie die Expertinnen und Experten unterschieden demgegenüber explizit zwischen gegenständlichen und mentalen Modellen und verstanden nur das Original als Gegenstand (Grosslight et al., 1991; Justi & Gilbert, 2003; Schwarz & White, 2005; Trier & Upmeier zu Belzen, 2009). Sie formulierten, dass Modelle Repräsentationen sind; einige sprechen neben dieser

Vorstellung vom Modell als Kopie (Justi & Gilbert, 2003). Zusammenfassend ist davon auszugehen, dass Schülerinnen und Schüler wie Lehrerinnen und Lehrer mehrheitlich eine große Ähnlichkeit zwischen Modell und Original annehmen und der subjektive, konstruktivistische Charakter von Modellen nicht im Fokus der Aufmerksamkeit steht.

Alternative Modelle zu einem Original

Eine weitere Kategorie von Vorstellungen bezieht sich auf alternative Modelle zu einem Original. Hier kann die Sichtweise eingenommen werden, dass es immer nur ein richtiges Modell gibt oder dass mehrere alternative Modelle existieren bzw. jedes Modell als Weiterentwicklung eines anderen Modells in einer historischen Abfolge steht (*uniqueness*, Justi & Gilbert, 2003). Mögliche Begründungen für die Existenz alternativer Modelle sind die Angepasstheit von Modellen an eine Zielgruppe (Lerntypen, Alter, Vorwissen etc.), verschiedene Blickwinkel in den Modellen, verschiedene Repräsentationsweisen oder Detailliertheitsgrade der Modelle, die Modellierung verschiedener Aspekte des Originals, verschiedene Ideen des Modellierers sowie konkurrierende Erklärungen eines Phänomens (*multiple models for the same thing*, Crawford & Cullin, 2005; *multiple models – to show different views of the same entity; two or more ideas or ways of explaining ‚it'; there are different ways to show or represent the same thing; you can make different models to test an entity*, Grosslight et al., 1991). Vorstellungen, die alternative Modelle mit verschiedenen Aspekten oder Detailliertheitsgraden der Modelle begründen, implizieren, dass bei der Modellierung Aspekte des Originals im Modell weggelassen oder hervorgehoben wurden (Grosslight et al., 1991). Diese Vorstellungen sind demnach mit einer Vorstellung von Modellen als vereinfachte Repräsentationen von Originalen verbunden. Wenn verschiedene Ideen des Modellierers als Grund für alternative Modelle genannt werden, verweist dies auf eine Vorstellung von Modellen als theoretische Rekonstruktionen von Originalen.

Die meisten Schülerinnen und Schüler sowie Lehrerinnen und Lehrer gingen davon aus, dass es alternative Modelle gibt. Bis zu ein Viertel nahm an, dass nur *ein* richtiges Modell existiert, das von der wissenschaftlichen *community*

evtl. noch nicht bestimmt wurde (Justi & Gilbert, 2003; Treagust et al., 2001, 2002; Trier & Upmeier zu Belzen, 2009; Van Driel & Verloop, 1999). Viele waren in ihren Antworten in dieser Kategorie unsicher und äußerten nur wenige Vorstellungen (Justi & Gilbert, 2003; Treagust et al., 2002). Die prominentesten Vorstellungen waren, dass alternative Modelle verschiedene Blickwinkel oder verschiedene inhaltliche Aspekte abbilden (Grosslight et al., 1991; Treagust et al., 2001, 2002; Trier & Upmeier zu Belzen, 2009). Die Schülerinnen und Schüler nahmen das Testen verschiedener Hypothesen mit alternativen Modellen kaum wahr. Sie thematisierten diesen Aspekt im Sinne eines Tests verschiedener Versionen eines Objekts, um zu sehen, ob das Modell „funktioniert" (Grosslight et al., 1991; Treagust et al., 2002). Die Expertinnen und Experten konzentrierten sich dagegen darauf, dass mit alternativen Modellen verschiedene Interessen bzw. Fragestellungen adressiert werden. Sie diskutierten die Kompatibilität verschiedener Modelle miteinander und thematisierten konkurrierende Modelle, die verschiedene Implikationen aufweisen oder verschiedene Vorhersagen treffen (Grosslight et al., 1991).

Zweck von Modellen

Vorstellungen zum Zweck von Modellen bilden eine vierte Kategorie. Die Schülerinnen und Schüler sowie Lehrerinnen und Lehrer nannten hierzu folgende Punkte: Modelle unterstützen die Kommunikation und können dazu dienen, Inhalte besser erklären zu können oder etwas – auch gedanklich – zugänglich oder verständlicher machen, sowie als Visualisierung den Denkprozess und die Kreativität unterstützen (*purpose of models*, Crawford & Cullin, 2005; *purpose of models – communication; observation; learning, understanding; accessibility or clarity*, Grosslight et al., 1991; *use*, Justi & Gilbert, 2003). Zu verschiedenen Kommunikationszwecken kann man dabei alternative Modelle einsetzen, so dass hier eine Verbindung zwischen diesen Kategorien von Vorstellungen besteht. Außerdem ist der Gedanke, dass Modelle etwas zugänglich machen, zum einen mit der Vorstellung verbunden, dass Modell und Original sich in Ort, Zeit, Größe, Blickwinkel oder Medien unterscheiden können. Zum anderen kann Zugänglichkeit im Sinne einer besseren Verständlichkeit und größeren Klarheit gegenüber dem Original dadurch erreicht werden,

dass Aspekte im Modell weggelassen oder hervorgehoben werden (Grosslight et al., 1991). Beides impliziert ein Verständnis vom Modell als vereinfachte Repräsentation.

Als weiterer Zweck wurde genannt, dass Modelle als Referenz genutzt werden können (*use*, Justi & Gilbert, 2003; *purpose of models – reference or example*, Grosslight et al., 1991). Eine weitere Einsatzmöglichkeit ist, bei Tests das Original durch das Modell zu ersetzen, um das Original zu schützen (*purpose of models*; Crawford & Cullin, 2005). Darüber hinaus können sie im Sinne des *model of modelling* (Justi & Gilbert, 2002) als Forschungswerkzeuge dem Testen von Vorhersagen dienen und zu Erkenntnissen über das Original führen (*purpose of models*, Crawford & Cullin, 2005, sowie Grosslight et al., 1991; *use* und *prediction*, Justi & Gilbert 2003).

Diese Vielfalt der Zwecke, mit denen Modelle eingesetzt werden, bildet sich in vielfältigen Vorstellungen hierzu ab. Zu dieser Kategorie wurden häufig mehrere Vorstellungen genannt (Grosslight et al., 1991; Justi & Gilbert, 2003), was auf die Differenziertheit der Vorstellungen hierzu schließen lässt. Die zentralen Vorstellungen zum Zweck von Modellen bezogen sich auf deren deskriptive Nutzung zur Visualisierung, Verbesserung der Verständlichkeit bzw. Kommunizierbarkeit und zur besseren Zugänglichkeit des Originals (Grosslight et al., 1991; Justi & Gilbert, 2003; Treagust et al., 2002; Trier & Upmeier zu Belzen, 2009; Van Driel & Verloop, 1999). AAAS-Items, die sich darauf beziehen, dass Modelle etwas visualisieren und Denkprozesse unterstützen, wurden von etwa drei Viertel der Schülerinnen und Schülern richtig beantwortet (AAAS, o. J.). Sie differenzierten dabei zwischen der Nutzung von Modellen in der Schule und in der Wissenschaft (Trier & Upmeier zu Belzen, 2009). Während jüngere Schülerinnen und Schüler die Visualisierbarkeit durch Modelle betonten, fokussierten auf ältere Verständlichkeit, Kommunizierbarkeit und Zugänglichkeit. Dieses Verständnis teilten auch die Expertinnen und Experten, bezogen es aber darauf, dass das durch Modelle vermittelte Verständnis eines Originals empirisch durch Beobachtungen am Modell bzw. Original geprüft werden kann (Grosslight et al., 1991). Dass mit Modellen Phänomene erklärt und Vorhersagen formuliert und getestet werden können, war für die Lehre-

rinnen und Lehrer evident (Justi & Gilbert, 2003), während viele Schülerinnen und Schüler hier unsicher waren und die wissenschaftliche Nutzung von Modellen in der Entwicklung von Ideen und Theorien nicht verstanden (Grosslight et al., 1991; Schwarz & White, 2005; Treagust et al., 2002).

Testen von Modellen

Vorstellungen zum Testen von Modellen können ebenfalls mehreren Kategorien zugeordnet werden. Eine bezieht sich darauf, dass Modelle als Prototyp fungieren und für Tests genutzt werden können, bevor etwas z. B. in Serie produziert wird. Modelle können geprüft werden, indem man eher ungerichtet ausprobiert, was passiert, das Verhalten des Modells in verschiedenen Situationen testet bzw. verschiedene Ideen über das Original mit dem Modell testet (*purpose of models – testing*, Grosslight et al., 1991). Dazu werden entweder das Modell mit dem Original oder Daten aus einem Modellexperiment mit Beobachtungen am Original verglichen (*validating/testing models*, Crawford & Cullin, 2005), was zu einer Revision des Modells führen kann (*purpose of models – testing*, Grosslight et al., 1991). Die Gültigkeit eines Modells wird vom Modellierer selbst, einer gesellschaftlichen Gruppe oder der *scientific community* geprüft (*validating/testing models*, Crawford & Cullin, 2005; *accreditation*, Justi & Gilbert, 2003).

Die Schülerinnen und Schüler verstanden das Testen von Modellen vor allem als ungerichtetes Ausprobieren, was passiert, oder als Testen eines Prototyps (Grosslight et al., 1991). Fast die Hälfte der befragten Schülerinnen und Schüler beantwortete das AAAS-Item hierzu entsprechend (AAAS, o. J.). Auch die Vorstellung, dass ein Modell durch das Parallelisieren mit dem Original überprüft wird, war verbreitet (Trier & Upmeier zu Belzen, 2009). Selten formulierten Schülerinnen und Schüler, dass mit Modellen Ideen getestet werden (Grosslight et al., 1991). Sie waren sich oft unsicher, ob Modelle der Formulierung und Prüfung von Vorhersagen dienen (Treagust et al., 2002).

Als Testkriterium forderten fast alle Schülerinnen und Schüler, dass Modelle genau und verständlich sind (Schwarz & White, 2005). Weder die Schülerin-

nen und Schüler noch eine Mehrheit der Lehrerinnen und Lehrer thematisierten, welche Instanz über die Gültigkeit von Modellen entscheidet (Justi & Gilbert, 2003). Die Lehrerinnen und Lehrer, die hierzu Stellung bezogen, sahen in der Regel den Modellierer als Instanz. Er testet die Gültigkeit eines Modells mithilfe empirischer Daten, die Modell und Theorie unterstützen, und prüft, inwiefern Ergebnisse mit dem Modell erklärbar sind. Die Unterstützung eines Modells durch die *Community* oder persönliche Gründe spielten eine untergeordnete Rolle in Vorstellungen zum Testen von Modellen (Justi & Gilbert, 2003; Treagust et al., 2001).

Veränderbarkeit von Modellen

Vorstellungen zur Veränderbarkeit von Modellen bilden eine weitere Gruppe von Kategorien. Manche Schülerinnen und Schüler sowie Lehrerinnen und Lehrer halten Modelle für unveränderlich (*changing a model*, Crawford & Cullin 2005; *changing a model – other or no*, Grosslight et al., 1991; *time*, Justi & Gilbert, 2003). Als Gründe für die Veränderung eines Modells führen sie die folgenden an:

- ästhetische Gründe (*changing a model - for aesthetic reasons;* Grosslight et al., 1991),
- struktureller oder funktioneller Fehler im Modell (*changing a model - something wrong with how the model was made*; Grosslight et al., 1991; *time*, Justi & Gilbert, 2003),
- mangelnde Übereinstimmung mit dem Original, z. B. aufgrund einer Veränderung des Originals, der Gewinnung neuer Erkenntnisse oder einer veränderten Vorstellung vom Original (*changing a model,* Crawford & Cullin, 2005; *changing a model - model not right because new information is found*; Grosslight et al., 1991),
- Anwendung des Modells in einem neuen Kontext, z. B. zum Testen neuer Ideen (*changing a model - model changes in accordance with how the purpose was matched;* Grosslight et al., 1991),
- mangelnde Übereinstimmung zwischen dem Verhalten des Modells und der Absicht des Modellierers (*changing a model,* Crawford & Cullin,

2005; *model changes in accordance with how the purpose was matched;* Grosslight et al., 1991),

- mangelnde Eignung des Modells, um Aspekte des Originals zu erklären (*changing a model,* Crawford & Cullin, 2005; *time*, Justi & Gilbert, 2003).

Grundsätzlich nahmen die meisten Schülerinnen und Schüler wie Lehrerinnen und Lehrer an, dass Modelle veränderbar sind (Grosslight et al., 1991; Justi & Gilbert, 2003; Treagust et al., 2001, 2002). Während jüngere Schülerinnen und Schüler diese Veränderbarkeit nur vage begründeten und dies auf die Veränderbarkeit des Originals zurückführten, sahen viele ältere Schülerinnen und Schüler neue Informationen über das Original als Ursache einer Veränderung des Modells (Grosslight et al., 1991; Trier & Upmeier zu Belzen, 2009). Viele befragte Schülerinnen und Schüler beantworteten die AAAS-Items, ob ein Modell aufgrund neuer Informationen über das Original verändert werden sollte, richtig (AAAS, o. J.). Weitere prominente Vorstellungen zur Veränderbarkeit von Modellen waren die, dass sie sich ändern, wenn das Modell fehlerhaft ist, Vorhersagen des Modells nicht zutreffen oder Aspekte des Originals mit einem Modell nicht erklärt werden können (Justi & Gilbert, 2003). Dagegen sahen die Schülerinnen und Schüler wie die Lehrerinnen und Lehrer das Modell selbst in der Regel nicht als Teil der Forschung (Justi & Gilbert, 2003; Grosslight et al., 1991). Die Expertinnen und Experten vertraten jedoch die Auffassung, dass Modelle geändert werden, um eine Interpretation der Welt zu unterstützen (Grosslight et al., 1991).

2.3.3 Komplexitätsgrade

Neben der Unterscheidung inhaltlicher Kategorien von Modellkompetenz werden unterschiedliche Komplexitätsgrade differenziert. Grosslight et al. (1991) graduieren Vorstellungen zu Modellen mithilfe empirisch bereits beschriebener Komplexitätsgrade von epistemologischen Sichtweisen auf Modelle und ihrer Nutzung in der Wissenschaft, d. h. mithilfe einer Graduierung im Wissenschaftsverständnis (*nature of science*). Dieses kann als Werte und Annahmen, die die Entwicklung wissenschaftlicher Erkenntnisse begleiten, beschrieben werden und somit als epistemologische Grundlage wissenschaftlicher Aktivitä-

ten (Lederman, 1992). Eine detaillierte Definition eines (Natur-)Wissenschaftsverständnisses ist Gegenstand eines andauernden Diskurses, da es sich mit grundsätzlich vorläufigen naturwissenschaftlichen Erkenntnissen auf einen dynamischen Gegenstand bezieht (Lederman, 2008). In den sogenannten *science wars* hat sich etwa das Verständnis von Wissenschaft als Entdeckung objektiver Realität zu einem sozial und kulturell geprägten Wissenschaftsbegriff gewandelt (Osborne, Collins, Ratcliffe, Millar & Duschl, 2003). Für das Wissenschaftsverständnis von Schülerinnen und Schülern sind aktuell folgende Punkte relevant (Lederman, 2008): Wissenschaftliche Erkenntnisse sind vorläufig, empirie-basiert, subjektiv, sozial und kulturell eingebettet sowie verbunden mit Kreativität. Außerdem differenzieren sie in Beobachtung und Erklärung sowie analog hierzu in Gesetz und Theorie.

Die auf dieser Grundlage formulierten Niveaus unterscheiden sich darin, wie die Beziehung von Modellen zur Realität und die Rolle, die Ideen mit Blick auf Modelle spielen, beschrieben werden. Niveau I entspricht einem naiv-realistischen Wissenschaftsverständnis, wonach Wissenschaft eine aktionale Produktion positiver Effekte bzw. ein rein kumulatives, deskriptives Faktensammeln ohne reflektierende Suche nach Zusammenhängen ist. Der Zweck von Wissenschaft ist demnach, Fragen zu beantworten sowie „auszuprobieren, ob etwas funktioniert" (Carey, Evans, Honda, Jay & Unger, 1989; Driver et al., 1996; Günther, Grygier, Kircher, Sodian, Thoermer, 2004). Theorie und Realität und somit Erklärung und Beschreibung werden hier gleichgesetzt, indem Wissen und Modelle als einfache wirklichkeitsgetreue Kopien der Realität gesehen werden (Carey et al., 1989; Driver et al., 1996; Grosslight et al., 1991).

Auf Niveau II wird Wissenschaft relativistisch als Suche nach experimentell überprüfbaren, induktiv beweisbaren Erklärungen verstanden (Carey et al., 1989; Driver et al., 1996; Grosslight et al., 1991; Günther et al., 2004). Sie hat diesem Verständnis nach zum einen das Ziel, Phänomene zu verstehen. Zum anderen zielt sie darauf, Zusammenhänge zwischen Phänomeneigenschaften zu finden, die auf Niveau II als lineare Kausalzusammenhänge interpretiert werden (Carey et al., 1989; Driver et al., 1996; Günther et al., 2004). Beschreibung und Erklärung werden dahingehend unterschieden, dass eine

Erklärung induktiv aus einer empirischen Beschreibung entwickelt werden kann. Dabei beziehen sich jedoch beide auf real existierende Eigenschaften (Carey et al., 1989; Driver et al., 1996). In Bezug auf Modelle werden deren Zweckgebundenheit und die subjektive Rolle des Modellierers wahrgenommen. Dabei steht jedoch das Modell bzw. die modellierte Realität statt der Modellbildung durch Wissenschaftlerinnen und Wissenschaftler im Vordergrund (Carey et al., 1989; Grosslight et al., 1991). Ein Modell muss diesem Verständnis nach nicht exakt dem Original entsprechen, sondern kann vereinfacht sein und Aspekte vernachlässigen bzw. hervorheben. Verschiedene Versionen von Modellen sind auf Niveau II denkbar; diese werden jedoch nicht so verstanden, dass sie verschiedene Ideen abbilden. Entsprechend wird ein Modell nicht mit Blick auf eine zugrunde liegende Idee getestet, sondern auf seine Funktionsfähigkeit (Driver et al., 1996; Grosslight et al., 1991).

Niveau III bildet ein konstruktivistisches Verständnis von Wissenschaft als Prozess ab, in dem zyklisch, kumulativ und vorläufig Wissen konstruiert wird (Carey et al., 1989; Driver et al., 1996; Grosslight et al., 1991). Für das Testen verschiedener vorläufiger Theorien im hypothetisch-deduktiven Verfahren werden Rahmentheorien sowie multiple Modelle benötigt, die dazu gezielt von einem Modellierer gebildet werden (Carey et al., 1989; Driver et al., 1996; Grosslight et al., 1991; Günther et al., 2004). Die Beziehung zwischen Beschreibung und Erklärung wird in dem Sinne problematisiert, dass eine Theorie nicht logisch aus der Beobachtung und Beschreibung eines Phänomens deduzierbar ist, da Erklärungen Vermutungen über theoretische Konstrukte beinhalten (Carey et al., 1989; Driver et al., 1996). Diesem Verständnis nach konstruiert der Modellierer Modelle, um Ideen zu entwickeln und zu testen und somit Erkenntnisse zu gewinnen. Die Informationen im zyklischen Prozess werden demnach über Modelle gewonnen (Grosslight et al., 1991).

Das Wissenschaftsverständnis unterscheidet sich nach Befunden von Urhahne, Kremer und Mayer (2008) mit den Jahrgangsstufen: Sie stellten bei Schülerinnen und Schülern fest, dass es in hohen Jahrgangsstufen elaborierter ist als in niedrigen. Analog kann man davon ausgehen, dass auch Modellkompetenz in höheren Jahrgangsstufen ausgeprägter ist. Schülerinnen und Schüler ver-

fügen in der Regel über ein Wissenschaftsverständnis, das sich von Niveau I zu II entwickelt (Carey et al., 1989; Driver et al., 1996; Grosslight et al., 1991; Sins et al., 2009). Expertinnen und Experten argumentieren dagegen auf Niveau III (Grosslight et al., 1991). Justi und Gilbert (2003) konnten jedoch die Niveaus nach Grosslight et al. (1991) für Lehrerinnen und Lehrer nicht replizieren. Sie fanden keine Muster, die diesen Niveaus entsprechen und verschiedene Aspekte von Vorstellungen zu Modellen umfassen. Sie schlussfolgern daraus, dass Lehrerinnen und Lehrer wahrscheinlich nicht über kohärente ontologische und epistemologische Sichtweisen verfügen. Insbesondere in den Kategorien *nature of models* und *use of models* formulierten die Lehrerinnen und Lehrer häufig mehrere parallele, unterschiedlich komplexe Vorstellungen.

Ein alternatives Niveaumodell formulieren Crawford und Cullin (2005). Sie nutzen die Kategorien von Grosslight et al. (1991) und nehmen darauf aufbauend eine Graduierung in vier Niveaus vor. Sie charakterisieren medial geprägte Vorstellungen auf Niveau I als *limited*, Vorstellungen zu Niveau II als *pre-scientific*, Niveau III als *emerging scientific* und Niveau IV als *scientific* (Tab. 1). Dabei formulieren sie für *purpose of models* und *validating/testing models* keine entsprechenden Vorstellungen auf Niveau I. Crawford und Cullin (2005) berichten, dass Lehramtsstudierende mehrheitlich über Vorstellungen im Bereich von Niveau II verfügen.

Tab. 1: Matrix von *Modelling Dimensions* (Crawford & Cullin, 2005, S. 316; Übersetzung und Kürzung E.T.).

	Limited – Niveau I	*Pre-Scientific* – Niveau II	*Emerging Scientific* – Niveau III	*Scientific* – Niveau IV
Purpose of Models	Lehr-Lernmittel zur Kommunizierbarkeit	visualisierendes Denkwerkzeug zur Formulierung einer Erklärung von Phänomenen	Ersatzobjekt für das Original in gefährlichen Tests	Forschungswerkzeug zur Gewinnung von Erkenntnissen über das Original
Designing and Creating Models	-	Einfluss der Ideen des Modellierers	Einbindung von Zusammenhängen innerhalb des Originals in das Modell	iterative Modellentwicklung orientiert an empirischen Daten
Changing a Model	keine Veränderbarkeit	neue Erkenntnisse	mangelnde Entsprechung des Verhaltens des Modells/Ideen des Modellierers	vorläufiger Charakter von Modellen, mangelnde Übereinstimmung mit Daten zum Original
Multiple Models for the same Thing	verschiedene Zielgruppen/ Lerntypen	verschiedene Ideen des Modellierers; verschiedene inhaltliche Aspekte	konkurrierende Modelle oder Theorien zur Erklärung eines Phänomens	verschiedene Annahmen oder verschiedene Fragestellungen zum Original
Validating/Testing Models	-	*scientific community* als externe Autorität	Vergleich des Verhaltens von Modell/Original	Vergleich experimenteller Daten des Modells mit Beobachtungen des Originals

Eine dritte Variante stammt von Stephens et al. (1999). Sie unterscheiden in Anlehnung an Driver et al. (1996) und Gentner (1989) vier Niveaus wissenschaftlicher Argumentation: Niveau I umfasst phänomenbasierte Argumente, die die Beschreibung eines Phänomens nicht klar von seiner Erklärung abgrenzen. Auf Niveau II basiert die Argumentation auf Zusammenhängen, die als Erklärung von Phänomenen induktiv aus den Daten heraus entwickelt wer-

den. Nur die beiden höchsten Niveaus beziehen sich auf Modelle, mit denen kausale Erklärungen entwickelt werden. Wenn sich die Argumentation auf Oberflächenmerkmale von Modellen bezieht, entspricht das nach Stephens et al. (1999) Niveau III. Die Einbeziehung von Zusammenhängen innerhalb des Modells entspricht Niveau IV. Laut Stephens et al. (1999) wird im Unterricht vor allem auf Niveau I und II argumentiert. Sie fanden demnach mit ihrer Niveau-Konzeption kaum modellbasierte Argumentationen.

Ein weiteres Niveaumodell stammt von Schwarz et al. (2009). Sie graduieren Vorstellungen zu Modellen in den Bereichen *scientific models as tools for predicting and explaning* und *models change as understanding improves* in jeweils vier Niveaus (Tab. 2). Jedes Niveau umfasst sowohl epistemologisches Metawissen über Modelle und die Modellbildung als auch eine Beschreibung der Performanz, d. h. den Umgang mit Modellen. Dabei gehen Schwarz et al. (2009) davon aus, dass epistemologische Vorstellungen den Umgang mit Modellen leiten, jeder Schritt im Umgang durch ähnliche epistemologische Vorstellungen gestützt ist und die Schritte interagieren. Dies geschieht z. B., wenn man bei der Entwicklung eines Modells mehrere mögliche Modelle evaluiert und sich dafür entscheidet, eine dieser Varianten zu realisieren. Den Umgang mit Modellen strukturieren sie in vier Schritte: die zielgerichtete Entwicklung von Modellen, die Nutzung von Modellen, den Vergleich und die Beurteilung von Modellen mit Blick auf deren Potenzial zur Vorhersage von Phänomenen sowie die Revision von Modellen auf der Grundlage von Daten. Die Niveaus, die sie auf dieser Grundlage formulieren, zeichnen sich durch eine Verschiebung des Fokus' von einer medialen zu einer methodischen Nutzung von Modellen aus (vgl. Tab. 2). Schwarz et al. (2009) beobachteten, dass Schülerinnen und Schüler sich vor allem auf Niveau 1 bis 2 bewegen, d. h. Modelle vor allem medial verstehen.

Tab. 2: *Learning progression* nach Schwartz et al. (2009, S. 9, 16; Übersetzung und Kürzung E.T.).

	Niveau 1	Niveau 2	Niveau 3	Niveau 4
scientific models as tools for predicting and explaning	Entwicklung und mediale Nutzung von Modellen als Illustration der direkt beobachtbaren Eigenschaften eines Phänomens	Entwicklung und Nutzung von Modellen zur Erklärung des eigenen Verständnisses von nicht direkt beobachtbaren Eigenschaften eines Phänomens	Entwicklung und Nutzung von Modellen als Denkwerkzeuge zur Erklärung und Vorhersage einzelner Aspekte von Phänomengruppen Einbeziehung und vergleichende Beurteilung alternativer Modelle	spontane Entwicklung und Nutzung von Modellen als Denkwerkzeuge zur Entwicklung und Beurteilung von Hypothesen über Phänomene
models change as understanding improves	Modell als unveränderlich und richtig oder falsch im Sinne einer guten oder schlechten Kopie eines Phänomens	Modifizierung von Details oder Verständlichkeit eines Modells aufgrund von Informationen einer Autorität (Lehrperson, Schulbuch etc.)	Verbesserung der Erklärungskraft eines Modells durch eine Optimierung der Passung zu Daten Vergleich alternativer Modelle mit Blick auf eine unterschiedliche Passung zu Daten	Berücksichtigung von Veränderungsmöglichkeiten eines Modells mit Blick auf dessen Erklärungskraft vor der Erhebung von Daten Einbeziehung alternativer Modelle für eine optimale Erklärungs-/Vorhersagekraft

Zusammenfassend lässt sich feststellen, dass Schülerinnen und Schüler eine eher mediale Perspektive auf Modelle haben. Diese wird im Vergleich zu einer methodischen Perspektive als weniger komplex eingeordnet. Klos, Henke, Kieren, Walpuski und Sumfleth (2008) sowie Koslowski (1996) berichten, dass

Mädchen über ein etwas elaborierteres Wissenschaftsverständnis verfügen als Jungen.

2.4 Modellkompetenz

Der psychologische geprägte Kompetenzbegriff nach Weinert (2001; vgl. Kapitel 2.1.1), wissenschaftstheoretische Konzepte zu Modellen (vgl. Kapitel 2.2) und empirische Befunde zur Strukturierung und Graduierung von Vorstellungen zu Modellen (vgl. Kapitel 2.3.2, 2.3.3) gehen in die theoriegeleitete Beschreibung von Modellkompetenz ein. Upmeier zu Belzen und Krüger (2010, S. 49) definieren diese wie folgt:

> „Modellkompetenz umfasst die Fähigkeiten, mit Modellen zweckbezogen Erkenntnisse gewinnen zu können und über Modelle mit Bezug auf ihren Zweck urteilen zu können, die Fähigkeiten, über den Prozess der Erkenntnisgewinnung durch Modelle und Modellierungen in der Biologie zu reflektieren sowie die Bereitschaft, diese Fähigkeiten in problemhaltigen Situationen anzuwenden."

Für die gezielte Förderung von Modellkompetenz wird ein Kompetenzmodell benötigt, um Ziele für den Umgang mit Modellen zu konkretisieren und Diagnosen einzuordnen. Hierzu gibt es mehrere Ansätze, die inhaltliche Bereiche und Niveaus von Modellkompetenz unterschiedlich differenziert beschreiben. Dieser Arbeit wird das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) zugrunde gelegt (Kapitel 2.4.1), das Modellkompetenz mit einem Fokus auf die Erkenntnisgewinnung beschreibt, inhaltliche Bereiche als Teilkompetenzen unterscheidet und sie in eine konzeptuelle und prozedurale Facette gruppiert sowie in Niveaus graduiert. Upmeier zu Belzen und Krüger (2010) decken damit einen Schlüsselbereich der Forschung zur Messung von Modellkompetenz ab (Kognitive Modellierung von Kompetenzen; Klieme & Leutner, 2006; Klieme et al., 2008; Koeppen et al., 2008). Weitere Schritte – die Entwicklung eines psychometrischen Modells, die Test- und Itemkonstruktion auf der Grundlage des Kompetenzmodells sowie Forschung zur Verwendung entsprechender Diagnoseinformationen (Klieme & Leutner, 2006; Klieme et al., 2008; Koeppen et al., 2008) – stehen noch aus.

Mit Blick auf ihre Domänenspezifität und auf die Abgrenzung zu verwandten Konstrukten wird Modellkompetenz in einem nomologischen Netzwerk beschrieben (Kapitel 2.4.2; vgl. Köller, 2008b). Dies legt die Grundlage, zur konvergenten bzw. diskriminanten Validierung Beziehungen zu anderen Konstrukten empirisch zu klären und somit Hypothesen zum Konstrukt Modellkompetenz zu prüfen.

2.4.1 Strukturierung und Graduierung

Verschiedene Ansätze zur Strukturierung und Graduierung von Fähigkeiten und Fertigkeiten im Umgang mit Modellen beziehen sich auf Hodsons Zielebenen für den naturwissenschaftlichen Unterricht (*learning science*, *doing science*, *learning about science*; Hodson, 1993; Kapitel 2.2.4). Henze et al. (2007) sehen den Schlüssel zur Erreichung dieser Ziele in einer zentralen Rolle von Modellen und Modellbildung im Unterricht. Im Rahmen eines *Public Understanding of Science* unterscheiden sie in Anlehnung an Hodson (1993) sowie Justi und Gilbert (2002) als Grobstruktur fachliche Inhalte (*the learning of scientific models*), Fähigkeiten (*the act of modelling*) und Wissenschaftsreflexion (*the critical reflection on the role and nature of models in science*). Dabei graduieren sie die Inhalte dieser Dimensionen nicht.

Meisert (2008, 2009) greift diesen Ansatz auf, differenziert Modellkompetenz analog in die drei Dimensionen ‚Modellwissen', ‚Modellverständnis' und ‚Modellarbeit' und ordnet diese verschiedenen Kompetenzbereichen zu. Das konzeptuelle ‚Modellwissen' verortet sie im Kompetenzbereich Fachwissen (KMK 2005), die prozedurale ‚Modellarbeit' im Kompetenzbereich Erkenntnisgewinnung (KMK, 2005) und das konzeptuelle ‚Modellverständnis' im übergeordneten Wissenschaftsverständnis. Sie beschreibt über den Ansatz von Henze et al. (2007) hinausgehend Wechselwirkungen zwischen den Dimensionen von Modellkompetenz. Dabei nimmt sie an, dass ‚Modellwissen' und ‚Modellarbeit' das ‚Modellverständnis' fundieren und erweitern, während dieses umgekehrt zur Ausprägung der Modellkompetenz in den beiden untergeordneten Dimensionen beiträgt. Das ‚Modellverständnis' graduiert Meisert (2009) nach Grosslight et al. (1991) in drei Niveaustufen und vier Teilaspekte (Modell-Original-Relation, Modellierer-Rolle, Entwicklungscharakter von Modellen, Funktion von

Modellen als Mittel der Erkenntnisgewinnung). Ihr Fokus liegt dabei jedoch weniger auf der Erkenntnisgewinnung mit Modellen als auf der Beziehung zwischen Modell und Original sowie der Wahrnehmung des Modellierers.

In einem dritten Ansatz beschreibt Leisner-Bodenthin (2006) deklaratives und prozedurales Wissen sowie den Selbständigkeitsgrad als Komponenten von Modellkompetenz im Physikunterricht (Abb. 4). Zum deklarativen Wissen, das ein Modellverständnis und Wissen zum Inhalt, zu bestimmten Annahmen und Idealisierungen von Modellen umfasst, formuliert sie einzelne Vorstellungen, die weder als Teilkompetenzen beschrieben noch in Niveaus graduiert werden. Darüber hinaus unterscheidet sie domänenspezifische und -übergreifende Modellkompetenz (vgl. auch Leisner & Mikelskis, 2004). Sie nimmt an, dass diese Facetten sich darin unterscheiden, inwiefern Schülerinnen und Schüler Wissen von bekannten auf unbekannte Kontexte transferieren können.

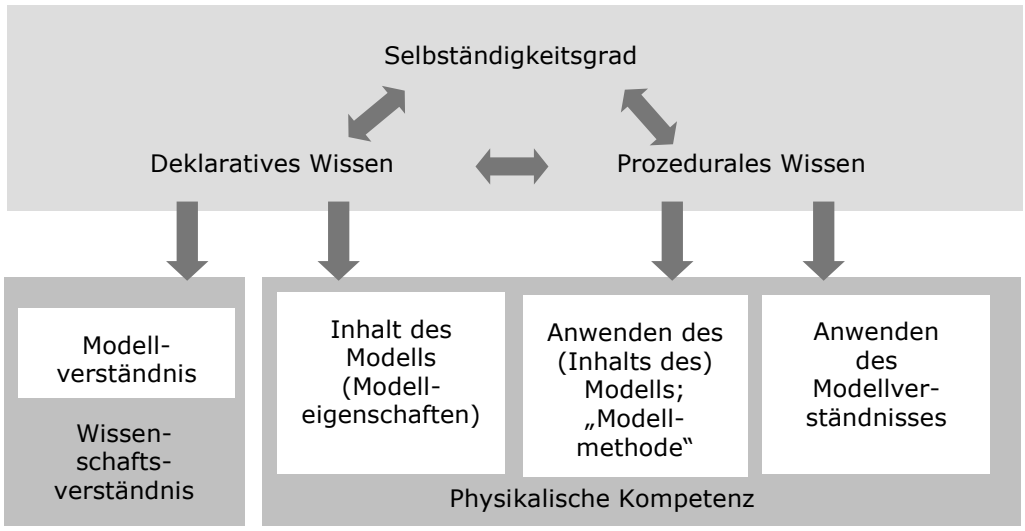


Abb. 4: Komponenten von Modellkompetenz nach Leisner-Bodenthin (2006).

Ein weiterer Ansatz zu einem Kompetenzmodell stammt von Crawford und Cullin (2005; vgl. Kapitel 2.3.3), die theoriebasiert eine Matrix mit fünf Aspekten und vier Niveaustufen entwickeln. Sie differenzieren nicht wie z. B. Meisert (2008) zwischen konzeptuellen und prozeduralen Facetten von Modellkompetenz. Darüber hinaus formulieren sie nicht für jede Teilkompetenz in

jedem Niveau Vorstellungen. Sie beschreiben die Niveaus über die Charakterisierung der einzelnen Bereiche, jedoch nicht übergreifend, d. h. unabhängig von den einzelnen Teilkompetenzen.

Schwarz et al. (2009) fokussieren in einem fünften Ansatz die Erkenntnisgewinnung mit Modellen (vgl. Kapitel 2.3.3). Dabei wollen sie klären, inwiefern Schülerinnen und Schüler Wissen über die Modellbildung vom spezifischen Kontext, in dem sie es erlernen, abstrahieren können. Hierfür beschreiben sie epistemologische Vorstellungen und Performanz integriert und beziehen Kommunikationsaspekte (*sensemaking* und *communicating understanding*) ein. Ein Nachteil dieser Konzeption ist, dass diese Anteile nicht getrennt voneinander erfasst werden können, so dass ein Förderbedarf weniger differenziert diagnostiziert werden kann. Studien, die sich auf einen psychologisch geprägten Kompetenzbegriff beziehen (vgl. Kapitel 2.1), stoßen bei dieser Strukturierung darüber hinaus auf die Schwierigkeit, dass die Beschreibung der Niveaus in den Dimensionen *scientific models as tools for predicting and explaning* und *models change as understanding improves* sich sowohl auf eine latente, als auch eine manifeste Ebene bezieht (vgl. Kapitel 2.3.3). Im Sinne des hier zugrunde gelegten Kompetenzbegriffs sollten diese Ebenen getrennt betrachtet und explizit aufeinander bezogen werden.

Die beschriebenen Ansätze führen Upmeier zu Belzen und Krüger (2010) zusammen. Ihr Kompetenzstruktur- und -niveaumodell zur Modellkompetenz im Kontext Biologieunterricht beinhaltet zwei große inhaltliche Bereiche: die eher deklarative Dimension ‚Kenntnisse über Modelle' und die eher prozedurale Dimension ‚Modellbildung'. Die Dimension ‚Kenntnisse über Modelle' beruht auf dem Wissenschaftsverständnis (Mayer, 2007), das unterschiedliche epistemologische und ontologische Positionen charakterisieren (Günther, 2006). Die Dimension ‚Modellbildung' beschreibt Fähigkeiten bei der Erkenntnisgewinnung mit Modellen im Modellbildungsprozess (Justi & Gilbert, 2002), die zum wissenschaftlichen Denken gehören (Mayer, 2007). Diese Fähigkeiten schließen die Reflexion des Modellbildungsprozesses sowie das Urteil über Modelle mit ein und führen zu einem allgemeinen Wissenschaftsverständnis (Treagust et al., 2002). Beide Dimensionen umfassen domänenübergreifendes

wie -spezifisches Wissen über Modelle (Leisner-Bodenthin, 2006), da sowohl abstrakte Konzepte über Modelle und die Modellbildung als auch konkrete Ideen zu spezifischen Modellen in die Definition der Dimensionen fallen. Upmeier zu Belzen und Krüger (2010) ordnen diesen Dimensionen aufbauend auf die empirischen Studien von Crawford und Cullin (2005), Justi und Gilbert (2002) sowie Grosslight et al. (1991) fünf Teilkompetenzen der Modellkompetenz zu (Tab. 3; vgl. Kapitel 2.3.2).

Tab. 3: Kompetenzmodell der Modellkompetenz (Upmeier zu Belzen & Krüger, 2010).

	Niveau I	Niveau II	Niveau III
Kenntnisse über Modelle			
Eigenschaften von Modellen	Modelle sind Kopien von etwas	Modelle sind idealisierte Repräsentationen von etwas	Modelle sind theoretische Rekonstruktionen von etwas
Alternative Modelle	Unterschiede zwischen den Modellobjekten	Ausgangsobjekt ermöglicht Herstellung verschiedener Modelle von etwas	Modelle für verschiedene Hypothesen
Modellbildung			
Zweck von Modellen	Modellobjekt zur Beschreibung von etwas einsetzen	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären	Zusammenhänge von Variablen für zukünftige neue Erkenntnisse voraussagen
Testen von Modellen	Modellobjekt überprüfen	Parallelisieren mit dem Ausgangsobjekt, Modell von etwas testen	Überprüfen von Hypothesen bei der Anwendung, Modell für etwas testen
Ändern von Modellen	Mängel am Modellobjekt beheben	Modell als Modell von etwas durch neue Erkenntnisse oder zusätzliche Perspektiven revidieren	Modell für etwas aufgrund falsifizierter Hypothesen revidieren

Die Teilkompetenzen differenzieren Upmeier und Krüger (2010) mit Bezugnahme auf Mahr (2008a) und aufbauend auf die Graduierung von Grosslight

et al. (1991) in drei Niveaus der Modellkompetenz. Diese beziehen sowohl eine mediale als auch eine methodische Perspektive auf Modelle ein und unterscheiden sich darin, welche Aspekte von Modellen in den Blick genommen werden: Auf einer relativ basalen Ebene wird das Modell als selbständiges Objekt (*Modellobjekt*, Mahr 2008a) betrachtet, ohne dass explizit Bezug auf ein Original genommen wird. Anspruchsvoller ist die Sichtweise, dass Modellen als *Modelle von etwas* Originale zugrunde liegen, die sie abbilden. Hier fehlt jedoch die Perspektive, dass die Anwendung von *Modellen für etwas* auch einen Rückschluss auf die modellierte Realität erlaubt und am Modell gewonnene Erkenntnisse auf das Original zurück übertragen werden können. Die Niveaus zeichnet demnach ein steigender Grad an Reflektiertheit aus (vgl. Bybee, 1997; Klieme, Avenarius et al., 2007; Klieme et al., 2008). Sie sind möglicherweise als Schritte beim Erwerb von Modellkompetenz interpretierbar (vgl. Klieme, Avenarius et al., 2007). Hierfür muss jedoch in einer längsschnittlichen Untersuchung geprüft werden, inwiefern die Niveaus empirisch differenziert werden können (s. hierzu Patzke & Upmeier zu Belzen, 2011).

In Bezug auf die ‚Eigenschaften von Modellen' können unterschiedliche Perspektiven auf die Beziehung zwischen Modell und Original eingenommen werden (*kinds of models*, Grosslight et al., 1991; *nature*, Justi & Gilbert, 2003). Modelle können als Kopien aufgefasst werden, die mit dem Original optimal übereinstimmen (Niveau I; Upmeier zu Belzen & Krüger, 2010). Das DNA-Modell von Watson und Crick stimmt etwa in der Struktur mit DNA überein, was zu der Vorstellung führen könnte, dass es eine Kopie der DNA ist. Der Gedanke, dass ein Modell eine Kopie des jeweiligen Originals ist und entsprechend stark mit ihm übereinstimmt, hängt damit zusammen, enge Parallelen zwischen Modell und Original zu ziehen (Testen von Modellen, Niveau II). Im Unterschied hierzu steht in der Teilkompetenz ‚Eigenschaften von Modellen' jedoch nicht die Frage nach der Einsetzbarkeit des Modells für einen bestimmten Zweck und somit die Anwendbarkeit des Modells im Vordergrund. Vielmehr ist der Fokus hier stärker deklarativ. In einer erweiterten Perspektive (Niveau II; Upmeier zu Belzen & Krüger, 2010) werden Modelle als idealisierte Repräsentationen wahrgenommen, wobei nur spezielle Aspekte des Originals

optimal und andere weniger gut repräsentiert werden (Stachowiak, 1973). Das DNA-Modell fokussiert als Stäbchenmodell etwa auf die Atombindungen. Anders als bei einem Kalottenmodell, wie dem von Pauling und Corey, steht die Atomgröße hier nicht im Vordergrund. Auf Niveau III werden Modelle als theoretische Rekonstruktionen von etwas angesehen (Upmeier zu Belzen & Krüger, 2010). Dabei wird wahrgenommen, dass die Eigenschaften des Modells durch den subjektiven Fokus des Modellierers bestimmt werden (Giere, 2009; Stachowiak, 1973). Watson und Crick ging es z. B. um den strukturellen Aufbau der DNA, den sie aus Röntgenbeugungsmustern rekonstruierten.

Für die Existenz ‚alternativer Modelle' sind verschiedene Gründe denkbar (*multiple models for the same thing*, Crawford & Cullin, 2005; *multiple models,* Grosslight et al., 1991; *uniqueness*, Justi & Gilbert, 2003). Wird nur das Modellobjekt betrachtet, werden Unterschiede zwischen verschiedenen Modellobjekten wie z. B. Größe, Farbe oder Material als Begründung für alternative Modelle angeführt (Niveau I; Upmeier zu Belzen & Krüger, 2010). Die DNA-Modelle von Watson und Crick sowie Pauling und Corey unterscheiden sich etwa darin, dass Stäbchen und Kugeln bzw. ausschließlich Kugeln zum Bau verwendet wurden. Wenn das Modell als Modell von etwas gesehen wird, werden alternative Modelle damit erklärt, dass verschiedene Aspekte des Originals abgebildet werden (Niveau II; Upmeier zu Belzen & Krüger, 2010). Während Watson und Crick sich auf die Atombindungen konzentrieren, fokussieren Pauling und Corey die Atomgrößen. Niveau III wird die Vorstellung zugeordnet, dass mit alternativen Modellen verschiedene Hypothesen über das Original repräsentiert werden (Upmeier zu Belzen & Krüger, 2010). Während dem DNA-Modell nach Watson und Crick z. B. die Hypothese zugrunde liegt, dass DNA die Struktur einer Doppelhelix aufweist, modellierten Pauling und Corey die Hypothese, dass DNA aus drei Helices besteht.

In der Teilkompetenz ‚Zweck von Modellen' (*purpose of models*, Crawford & Cullin, 2005; *purpose of models,* Grosslight et al., 1991; *use; prediction*, Justi & Gilbert, 2003) wird auf Niveau I der Zweck eines Modells darin gesehen, etwas zu veranschaulichen (Upmeier zu Belzen & Krüger, 2010). Das DNA-Modell von Watson und Crick kann etwa zur Veranschaulichung der Anordnung

der Bausteine der DNA genutzt werden. Als Modell *von* etwas erklärt ein Modell einen Zusammenhang im Original (Niveau II; Upmeier zu Belzen & Krüger, 2010) wie etwa das Zahlenverhältnis der Basen. Als Modell *für* etwas wird ein Modell für Voraussagen über das Original verwendet (Niveau III; Upmeier zu Belzen & Krüger, 2010), z. B. über Replikationsmechanismen der DNA.

In der Teilkompetenz ‚Testen von Modellen' wird deutlich, auf welche Weise Modelle geprüft werden (*validating/testing models*, Crawford & Cullin, 2005; *purpose of models – testing*, Grosslight et al., 1991). Wird nur das Modellobjekt in den Blick genommen, handelt es sich lediglich um eine strukturelle oder funktionelle Prüfung des Modellobjekts bei der z. B. das Material auf seine Widerstandsfähigkeit überprüft wird (Niveau I; Upmeier zu Belzen & Krüger, 2010). Es ist z. B. denkbar, dass das Material des DNA-Modells nicht stabil genug gewesen wäre, um das Modell standsicher aufzubauen. Bei theoretischen Modellen (z. B. Stoffkreisläufen) werden auf diesem Niveau Darstellungsfehler in grafischen Visualisierungen identifiziert. Eine weitere Perspektive ist das Parallelisieren von Modell und Original, bei dem die Passung mit der modellierten Realität (Giere, 2001, 2004) geprüft wird (Niveau II; Upmeier zu Belzen & Krüger, 2010). Diese Prüfung basiert auf bereits bestehenden Kenntnissen über das Original. DNA-Modelle, die keine helikale Struktur aufwiesen, wurden etwa aufgrund der mangelnden Übereinstimmung mit Röntgenbeugungsmustern verworfen. Bei der Erkenntnisgewinnung mit Modellen werden Hypothesen über das Original mit dem Modell überprüft (Giere et al., 2006), die entweder neue Erkenntnisse über die repräsentierte Realität generieren oder zur Änderung des Modells führen können (Niveau III; Upmeier zu Belzen & Krüger, 2010). DNA-Modelle, die andere Basenpaarungen als Adenin (A) – Thymin (T) und Guanin (G) – Cytosin (C) beinhalteten, waren z. B. aufgrund der räumlichen Struktur der Bindungen instabil, so dass diese Paarungen ausgeschlossen und die Modelle verworfen wurden.

Es können unterschiedliche Gründe angeführt werden, warum ein Modell verändert werden muss (Teilkompetenz ‚Ändern von Modellen'; *changing a model*, Crawford & Cullin, 2005, sowie Grosslight et al., 1991; *time*, Justi & Gilbert 2003): Dies kann 1. ein grundlegender Fehler im Modellobjekt sein, z. B.

im Material oder in Bezug auf die Funktionalität (etwa ein instabiles DNA-Modell; Niveau I; Upmeier zu Belzen & Krüger, 2010), 2. ein neuer, nicht im Modell berücksichtigter Erkenntnisstand, z. B. aufgrund neuer Informationen über das Original (etwa durch neue Techniken zur Strukturaufklärung wie beispielsweise zum Röntgen von DNA; Niveau II; Upmeier zu Belzen & Krüger, 2010), oder 3. eine falsifizierte Hypothese (Giere et al., 2006) als Grundlage des Modells (etwa das Tripelhelix-DNA-Modell von Pauling und Corey; Niveau III; Upmeier zu Belzen & Krüger, 2010).

2.4.2 Beziehungen zu anderen Konstrukten

Die theoriegeleitete Beschreibung von Modellkompetenz umfasst neben ihrer Strukturierung und Graduierung die Beziehungen zu anderen Konstrukten. Modellkompetenz lässt sich in das Rahmenkonzept wissenschaftsmethodischer Kompetenzen nach Mayer (2007) eingliedern (Tab. 1): ‚Kenntnisse über Modelle' zählen zum Wissenschaftsverständnis, ‚Modellbildung' zum wissenschaftlichen Denken sowie zu manuellen Fertigkeiten (Upmeier zu Belzen & Krüger, 2010). Somit ist Modellkompetenz in den Kompetenzbereich Erkenntnisgewinnung (KMK, 2005) einzuordnen und vom biologischen Fachwissen abgrenzbar. Da sie für die Domäne Biologie definiert ist, ist anzunehmen, dass sie vor allem mit Leistungen in naturwissenschaftlichen Fächern zusammenhängt.

Das Wissenschaftsverständnis ist neben Mahrs (2008a) Konzeption eines dynamischen Modellbegriffs Grundlage der Graduierung von Modellkompetenz in Niveaus (vgl. Kapitel 2.3.3). Deshalb ist Modellkompetenz mit diesem Konstrukt besonders eng verbunden. Driver et al. (1996) nennen das höchste Niveau wissenschaftlicher Argumentationsweisen *model-based reasoning*. Demnach werden Erklärungen für Phänomene nicht logisch aus Daten abgeleitet, sondern modelliert (vgl. Kapitel 2.3.3). Sie beziehen sich damit auf die Schlussfolgerung vom Modell als Original und somit nach Mahr (2008a) ausschließlich auf das Modell *für* etwas (vgl. Kapitel 2.2.3). Demnach ist *model-based reasoning* ein Teilbereich von Modellkompetenz.

Kremer, Grube, Urhahne und Mayer (2010) fanden Zusammenhänge zwischen dem Wissenschaftsverständnis und wissenschaftlichem Denken in mittlerer Höhe. Mit dem wissenschaftlichen Denken ist die ‚Modellbildung' auf Niveau III verbunden, das sich auf die Erkenntnisgewinnung mit Modellen bezieht. Es ist zu vermuten, dass Schülerinnen und Schüler beim experimentellen Testen und Ändern von Modellen im hypothetisch-deduktiven Verfahren ähnliche Schwierigkeiten wie beim Experimentieren mit Originalen haben. Nach Grube (2011) und Wellnitz (2012) ist zu erwarten, dass diese Schwierigkeiten beide Geschlechter gleichermaßen betreffen. Wie Studien belegen, kann methodisches Wissen über Ziele und Vorgehensweisen beim Experimentieren nicht vorausgesetzt werden. Bereits bei der Formulierung von Hypothesen grenzen Schülerinnen und Schüler Inhalte häufig zu stark ein (*positive capture*, Hammann, Phan, Ehmer & Bayrhuber, 2006). Sie versuchen bei der Planung und Durchführung eines Experiments in der Regel einen erwarteten Effekt bzw. ein besonders „gutes" Ergebnis zu erzielen statt systematisch nach Kausalzusammenhängen zu suchen. Dabei wählen sie häufig keinen aussagefähigen Kontrollansatz aus und variieren die relevanten Variablen unsystematisch. Die Tendenz, nur bestätigende Daten zu erheben und als Evidenzen für die vermuteten Zusammenhänge zu werten, wird als Strategie des positiven Testens, *confirmation bias* oder *failure to seek disconfirmation* bezeichnet. Inhaltliche Eingangsüberzeugungen beeinflussen häufig stark die Schlussfolgerungen aus den Ergebnissen (Chinn & Brewer, 1993; Hammann et al., 2006).

Ein weiteres Konstrukt, mit dem Modellkompetenz in Verbindung steht, betrifft den Repräsentationscharakter von Modellen. Dieser berührt Fähigkeiten, mit Repräsentationen umzugehen, und zwar 1. zu erkennen, welche Repräsentationstypen sinnvoll genutzt werden können, 2. den Charakter (Analogien zum Original) und Modus der Repräsentation zu identifizieren, 3. die Funktion der Repräsentation zu erkennen sowie 4. Inhalte zwischen verschiedenen Repräsentationen übertragen zu können (Chittleborough & Treagust, 2007; vgl. Ainsworth, 2008). Kozma und Russell (1997) bezeichnen die zielgerichtete, begründete Auswahl, Konstruktion und Nutzung von Repräsentationen sowie den Transfer von Inhalten zwischen verschiedenen Repräsentationsformen als *representational competence*. Diese bezieht sich somit vor allem auf den Um-

gang mit Oberflächenmerkmalen von Modellen und das Erfassen des Modellinhalts, d. h. sie fokussiert basale Anteile der medialen Rolle von Modellen, und kann deshalb als Grundlage für Modellkompetenz angesehen werden. Da Diagramme depiktionale, abstrakte Repräsentationen sind, die wie andere Modelle die Struktur des dargestellten Sachverhalts (des Originals) erhalten (vgl. Kapitel 2.2.2), ist Diagrammkompetenz ein Teilbereich von Modellkompetenz. Lachmayer (2008) beschreibt als Dimensionen von Diagrammkompetenz die Informationsentnahme, die Konstruktion von Diagrammen und die Integration von Informationen aus Diagrammen und sprachlichen Repräsentationen. Damit ist Diagrammkompetenz analog zur *representational competence* definiert und nimmt ebenfalls vor allem das Modellobjekt sowie das Erfassen des Modellinhalts in den Blick.

Da Modelle sich auf zusammenhängende Phänomeneigenschaften beziehen (Buckley & Boulter, 2000; Johnson-Laird, 1983; Vosniadou, 2002; vgl. Kapitel 2.2.1), weisen sie Systemcharakter auf. Aus diesem Grund stellt sich die Frage nach dem Verhältnis von Modell- zu Systemkompetenz bzw. Systemdenken. Systemkompetenz definiert Sommer (2006) als Fähigkeiten zum Erkennen, Beschreiben, Verstehen, Darstellen und Umgehen mit Systemeigenschaften. Dies umfasst die Dimension ‚strukturelles Systemdenken', die das Verständnis von Systemen beinhaltet, sowie die Dimension ‚prozedurales Systemdenken', die die Entwicklung von Modellen berührt (Brandstädter, Harms & Großschedl, 2012): Um Systemelemente und Beziehungen zwischen diesen Elementen zu identifizieren und vorherzusagen, muss ein mentales sowie ein konzeptuelles Modell gebildet werden. Modellkompetenz ist somit eng mit Systemkompetenz verzahnt. Für spezifische Darstellungsformen von dynamischen Systemen wie z. B. verbale Beschreibungen oder Flussdiagramme ist Modellkompetenz eine Voraussetzung von Systemkompetenz (Sommer, 2006). Das Konstrukt der Systemkompetenz richtet jedoch eher einen holistischen Blick auf fachliche Inhalte, während Modellkompetenz die Intentionalität von Modellen und ihre Rolle in der Erkenntnisgewinnung fokussiert.

Da ein grundlegendes Merkmal von Kompetenzen ihre Erlernbarkeit ist, die sie von allgemeinen kognitiven Fähigkeiten abgrenzt, ist es wichtig, bei der Kon-

struktion von Kompetenzmodellen die Beziehung der jeweiligen Kompetenz zu allgemeinen kognitiven Fähigkeiten zu berücksichtigen und empirisch zu prüfen (Klieme & Leutner, 2006; Koeppen et al., 2008). Typischerweise ist dieser Zusammenhang recht hoch, mit naturwissenschaftlichen Kompetenzen in PISA 2003 $r = .68$ (Leutner, Klieme, Meyer & Wirth, 2004; vgl. Hartig & Klieme, 2006). Chittleborough und Treagust (2007) argumentieren, dass dieser Zusammenhang für Modellkompetenz (*modelling ability*) relevant ist.

2.5 Problemstellung

Es gibt verschiedene Ansätze, Vorstellungen über Modelle zu strukturieren (u. a. Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; vgl. Kapitel 2.3.2), zu graduieren (Crawford & Cullin, 2005; Grosslight et al., 1991; Schwarz et al., 2009; Stephens et al., 2009; vgl. Kapitel 2.3.3) und ein Kompetenzmodell zu formulieren (u. a. Crawford & Cullin, 2005; Leisner-Bodenthin, 2006; Meisert, 2008; Schwarz et al., 2009; vgl. Kapitel 2.4.1). Dennoch liegt bislang kein Kompetenzmodell der Modellkompetenz vor, das empirisch überprüft wurde.

Das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) nimmt theoriegeleitet sowohl eine Strukturierung als auch Graduierung von Modellkompetenz vor, unterscheidet darin deklarative und prozedurale Anteile und ist ausschließlich einer latenten Ebene zuzuordnen. Außerdem weist es einen engen Bezug zum Kompetenzbereich Erkenntnisgewinnung der Bildungsstandards (KMK, 2005) auf und ist somit an diese normativen Vorgaben anschlussfähig. Deshalb wird es als Grundlage für die vorliegende Untersuchung herangezogen, die auf seine empirische Überprüfung abzielt. Dabei werden zwei Ziele verfolgt: die Operationalisierung des Kompetenzmodells in MC-Items und die empirische Beschreibung von Modellkompetenz mithilfe dieser Items.

Operationalisierung des Kompetenzmodells

Die Aussagekraft von Daten zur empirischen Überprüfung von Kompetenzmodellen hängt von der Validität und somit von der Qualität der Operationalisierung der entsprechenden Kompetenz in Items ab. Entsprechend stellt sich die

Frage, welche empirischen Indizien es dafür gibt, dass die Operationalisierung von Modellkompetenz in MC-Items die Erfassung von Modellkompetenz ermöglicht. In Bezug auf die Validität der entwickelten MC-Items ist die Legitimität von zwei Schlüssen relevant (vgl. Hartig, Frey & Jude, 2012): Zum einen sollen die Items das Kompetenzmodell adäquat repräsentieren; zum anderen soll die Bearbeitung der einzelnen Items als Indikator der entsprechenden Kompetenz interpretierbar sein. Die Legitimität dieses Repräsentations- und Interpretationsschlusses ist die Voraussetzung für die empirische Überprüfung des Kompetenzmodells (Upmeier zu Belzen & Krüger, 2010) mit diesen Items und muss somit bei der Testentwicklung geprüft werden.

Empirische Beschreibung von Modellkompetenz

Upmeier zu Belzen und Krüger (2010) differenzieren Modellkompetenz in zwei Dimensionen, ‚Kenntnisse über Modelle' sowie ‚Modellbildung', denen insgesamt fünf Teilkompetenzen, ‚Eigenschaften von Modellen', ‚Alternative Modelle', ‚Zweck', ‚Testen' und ‚Ändern von Modellen', zugeordnet werden. Diese Vorstellungen werden in drei Niveaus graduiert (vgl. Kapitel 2.4.1). Eine Fragestellung des hier vorgestellten Projekts bezieht sich auf die empirische Abbildung dieser Struktur.

> Inwiefern bildet sich die theoriegeleitete Struktur des Kompetenzmodells in einer empirischen Datenstruktur ab?

> Aufbauend auf die empirischen Studien zu Vorstellungen zu Modellen (u. a. Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; vgl. Kapitel 2.3.2) und deren Strukturierung in ein Kompetenzmodell (Upmeier zu Belzen & Krüger, 2010; vgl. Kapitel 2.4.1) ist anzunehmen, dass sich fünf Teilkompetenzen von Modellkompetenz unterscheiden lassen. Eine weitere aus der Theorie ableitbare Möglichkeit ist, dass eine deklarative und eine prozedurale Dimension unterschieden werden können (‚Kenntnisse über Modelle' und ‚Modellbildung', Upmeier zu Belzen & Krüger, 2010; vgl. Kapitel 2.4.1). Entsprechend werden drei Hypothesen zur **Strukturierung** von Modellkompetenz aufgestellt:

H0: Modellkompetenz lässt sich am besten durch ein einzelnes Kontinuum abbilden. Alle Items laden auf einen globalen Faktor, so dass sich keine Dimensionen unterscheiden lassen (eindimensionale Struktur von Modellkompetenz, Abb. 5), d. h. empirisch weist das eindimensionale Modell eine bessere Passung als das zwei- bzw. fünfdimensionale Modell auf.

H1: Modellkompetenz lässt sich am besten durch die Dimensionen ‚Kenntnisse über Modelle' sowie ‚Modellbildung' abbilden (zweidimensionale Struktur von Modellkompetenz, Abb. 5), d. h. empirisch weist das zweidimensionale Modell eine bessere Passung als das ein- bzw. fünfdimensionale Modell auf und die Dimensionen korrelieren im schwachen bis mittleren Bereich.

H2: Modellkompetenz lässt sich am besten durch die Teilkompetenzen ‚Eigenschaften von Modellen', ‚Alternative Modelle', ‚Zweck', ‚Testen' und ‚Ändern von Modellen' abbilden (fünfdimensionale Struktur von Modellkompetenz, Abb. 5), d. h. empirisch weist das fünfdimensionale Modell eine bessere Passung als das ein- bzw. zweidimensionale Modell auf und die Teilkompetenzen korrelieren im schwachen bis mittleren Bereich.

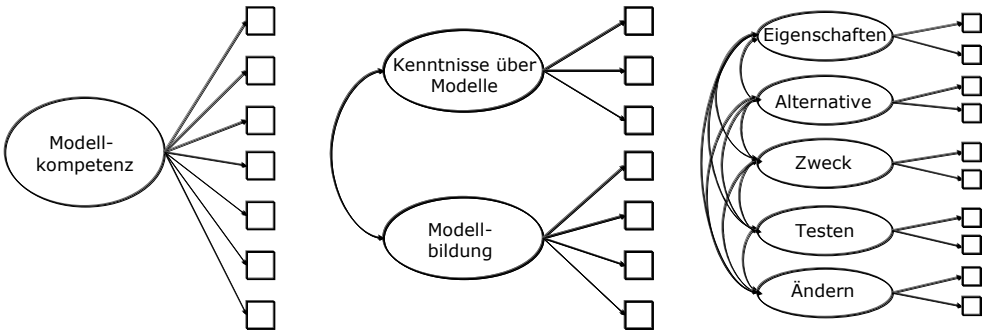


Abb. 5: Ein- (H0), zwei- (H1) bzw. fünfdimensionales (H2) Strukturmodell von Modellkompetenz.

Nach empirischen Untersuchungen (Crawford & Cullin, 2005; Grosslight et al., 1991; vgl. Kapitel 2.3.3) und aufbauend auf den dynamischen Modellbegriff nach Mahr (2008) unterscheiden Upmeier zu Belzen und Krüger (2010) drei Niveaus von Modellkompetenz (vgl. Kapitel 2.4.1). Entsprechend ist anzunehmen, dass diese **Graduierung** sich empirisch in Itemschwierigkeiten abbildet:

> **H3:** Die Niveaus I bis III bilden ansteigende Anforderungen an Schülerinnen und Schüler ab, d. h. die mittleren Itemschwierigkeiten steigen mit den a priori zugeordneten Niveaus von Modellkompetenz an und die a priori Zuordnung der Items zu den Niveaus erklärt substanzielle Anteile der Varianz der Itemschwierigkeit.

Modellkompetenz ist dem hier zugrunde gelegten Kompetenzbegriff nach als erlernbar definiert (z. B. Klieme & Leutner, 2006; Weinert, 2001; vgl. Kapitel 2.1.1, 2.4). Da Modelle im Biologieunterricht eingesetzt werden und deshalb Lerngelegenheiten vorhanden sind, ist anzunehmen, dass die Ausprägung von Modellkompetenz mit der Jahrgangsstufe steigt. Auch der Befund von Urhahne et al. (2008), dass das Wissenschaftsverständnis in höheren Jahrgangsstufen ausgeprägter ist, und der enge Bezug dieses Konstrukts zu Modellkompetenz sprechen für diese Annahme. Für das biologische Fachwissen wurde ein Leistungsanstieg auf die Jahrgangsstufe und nicht das Alter zurückgeführt (Schmiemann, 2010), so dass zu vermuten ist, dass dies auch für Modellkompetenz der Fall ist. Mit Blick auf die **Erlernbarkeit** von Modellkompetenz stellt sich demnach zunächst folgende Frage:

> Inwiefern unterscheidet sich in den Jahrgangsstufen 7 bis 10 die Performanz der Schülerinnen und Schüler mit Blick auf Modellkompetenz?

> **H4:** Die Ausprägung von Modellkompetenz ist von der Jahrgangsstufe abhängig, d. h. in höheren Jahrgangsstufen werden höhere Leistungen erzielt.

> **H5:** Das Alter hat über die Jahrgangsstufe hinaus keinen Einfluss auf die Modellkompetenz, d. h. unter Kontrolle der Jahrgangsstufe hat es keinen eigenen Effekt auf die Personenfähigkeiten.

Für eine empirische Beschreibung von Modellkompetenz sind außerdem ihre Beziehungen zu anderen Konstrukten und Variablen relevant (Köller, 2008a). Hierzu werden mit Blick auf die **Domänenspezifität** und die **Erlernbarkeit** von Modellkompetenz Schulnoten sowie allgemeine kognitive Fähigkeiten in den Blick genommen.

Inwiefern ist Modellkompetenz erlernbar und domänenspezifisch?

Da angenommen wird, dass Modellkompetenz als Kompetenz erlernbar ist (vgl. Kapitel 2.1) und allgemeine kognitive Fähigkeiten im Vergleich dazu stabiler sind, ist sie von allgemeinen kognitiven Fähigkeiten abzugrenzen, auch wenn sie mit ihnen in Verbindung steht (Chittleborough & Treagust, 2007; vgl. Koeppen et al., 2008). Spezifische Kompetenzen hängen typischerweise eng mit allgemeinen kognitiven Fähigkeiten zusammen, in PISA 2003 korrelierten sie mit naturwissenschaftlicher Kompetenz z. B. in Höhe von $r = .68$ (Leutner et al., 2004). Dies ist demnach auch für Modellkompetenz anzunehmen:

> **H6:** Modellkompetenz und allgemeine kognitive Fähigkeiten hängen stark miteinander zusammen, sind aber voneinander abgrenzbar, d. h. Personenfähigkeiten im Bereich Modellkompetenz korrelieren stark mit allgemeinen kognitiven Fähigkeiten.

Modellkompetenz wird als biologiespezifische Kompetenz betrachtet, die in das Rahmenkonzept (natur-)wissenschaftsmethodischer Kompetenzen nach Mayer (2007) eingegliedert werden kann (vgl. Kapitel 2.4.2). Nach Trautwein, Lüdtke, Becker, Neumann und Nagy (2008) kann ein moderater Zusammenhang zwischen Schulnoten und Lernleistungen angenommen werden, so dass Schulnoten als Schätzung für Lernleistungen in unterschiedlichen Fächern grundsätzlich geeignet sind. Für das wissenschaftliche Denken stellte Grube (2011) einen Zusammenhang von $r = -.13$ fest. Wellnitz (2012) fand für na-

turwissenschaftliche Untersuchungen einen Zusammenhang von r = -.25. Auch in PISA wurden Zusammenhänge zwischen naturwissenschaftlicher Kompetenz und Schulnoten in naturwissenschaftlichen Fächern diagnostiziert ($r_{Biologie}$ = -.36; r_{Physik} = -.34; r_{Chemie} = -.35; Schütte, Frenzel, Asseburg & Pekrun, 2007). Urhahne et al. (2008) berichten außerdem einen Zusammenhang von -.11 ≤ r ≤ -.23 zwischen Facetten des Wissenschaftsverständnisses und Lernleistungen in den naturwissenschaftlichen Fächern. Da Modellkompetenz mit diesem Konstrukt in Verbindung steht und hier für den Bereich der Biologie untersucht wird, sind auch hierfür entsprechende Zusammenhänge anzunehmen. Es ergibt sich folgende Hypothese:

> **H7:** Modellkompetenz hängt enger mit Leistungen in naturwissenschaftlichen Fächern zusammen als mit Leistungen in sprachlichen Fächern, d. h. die Personenfähigkeiten im Bereich Modellkompetenz korrelieren in mittlerer Höhe mit Schulnoten in naturwissenschaftlichen Fächern, insbesondere Biologie, und in geringerer Höhe mit Schulnoten in Deutsch und der ersten Fremdsprache.

Nach Studien von Klos et al. (2008) sowie Koslowski (1996) verfügen Mädchen über etwas höhere wissenschaftsmethodische Kompetenzen als Jungen. Prenzel, Schöps et al. (2007) berichten einen leichten, jedoch nicht signifikanten Vorsprung der Jungen und eine Unterrepräsentierung der Mädchen in der Spitzengruppe von PISA 2006. Nach Grube (2011) und Wellnitz (2012) besteht jedoch kein Geschlechtereffekt beim wissenschaftliche Denken bzw. bei Kompetenzen für naturwissenschaftliche Untersuchungen. Für Modellkompetenz ist demnach kein großer Geschlechterunterschied zu erwarten. Der enge theoretische Zusammenhang, in dem Modellkompetenz mit dem Wissenschaftsverständnis steht (Kapitel 2.3.3), bildet sich vermutlich auch empirisch ab.

> **H8:** Mädchen verfügen über eine ähnlich entwickelte Modellkompetenz als Jungen, d. h. die Personenfähigkeiten im Bereich Modellkompetenz unterscheiden sich nicht zwischen den Geschlechtern.

H9: Modellkompetenz hängt mit Wissenschaftsverständnis zusammen, unterscheidet sich aber von ihm, d. h. Personenfähigkeiten im Bereich Modellkompetenz korrelieren in mittlerer Höhe mit Wissenschaftsverständnis.

Im hier vorgestellten Projekt soll Modellkompetenz mit einem kontextgebundenen Instrument mit geschlossenem Antwortformat erhoben werden (Kapitel 2.3.1, vgl. bik-Projekte in Kapitel 2.1.2). Da Aufgaben in geschlossenem Antwortformat vergleichsweise viel Text enthalten, der für eine adäquate Itembearbeitung gelesen und verstanden werden muss, ist ein starker Zusammenhang zu Lesefähigkeiten zu erwarten. In PISA 2003 (Leutner et al., 2004) korrelierten Fähigkeiten im Bereich Naturwissenschaften z. B. in Höhe von $r = .83$ mit Lesefähigkeiten.

H10: Die Bearbeitung eines Modellkompetenz-Tests mit geschlossenem Antwortformat hängt stark mit Lesefähigkeiten zusammen, wird aber nicht vollständig durch sie erklärt, d. h. Personenfähigkeiten im Bereich Modellkompetenz korrelieren stark mit Lesefähigkeiten.

3 Operationalisierung des Kompetenzmodells

Aufbauend auf die Kompetenzmodellierung, in der Strukturen und Niveaus sowie die mögliche Kompetenzentwicklung beschrieben werden, sollten Kompetenzmodelle als Struktur-, Niveau- sowie Entwicklungsmodelle empirisch überprüft werden (vgl. Kapitel 2.1.1). Dafür wird ein Test benötigt, von dessen Bearbeitung auf die relevante Kompetenz geschlossen werden kann (Klieme et al., 2008; Koeppen et al., 2008; Abb. 6).

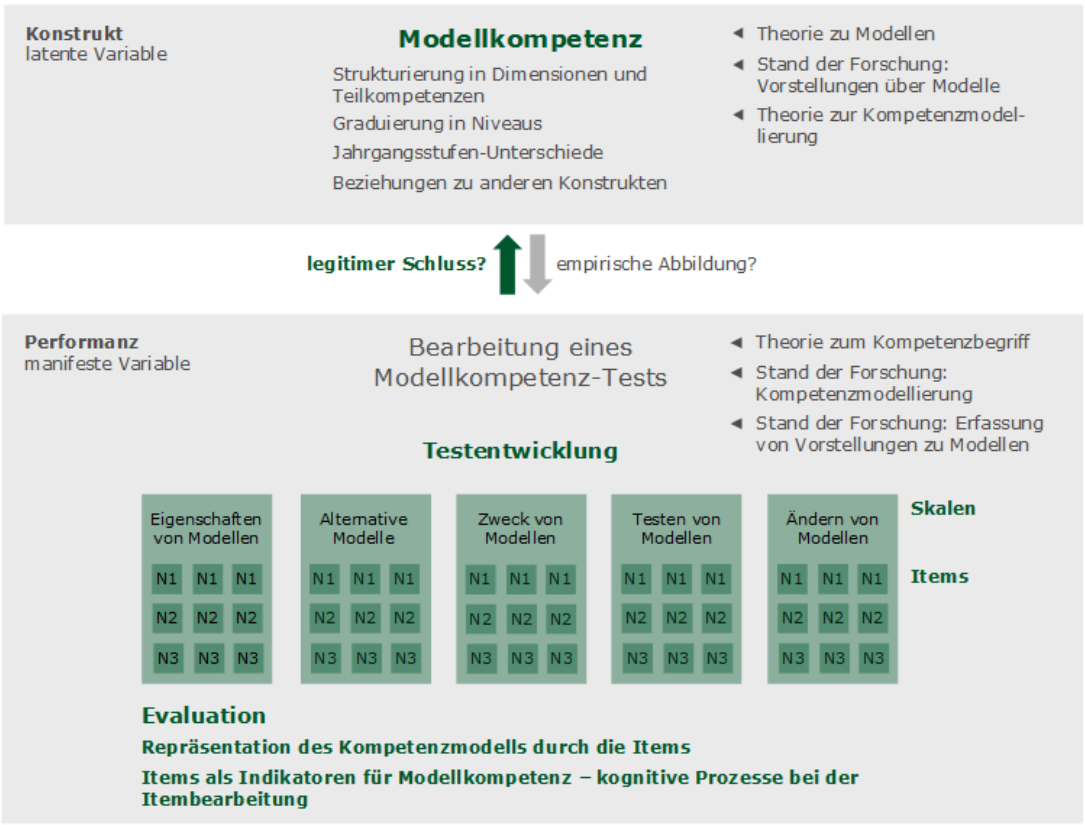


Abb. 6: Konzeption des Projekts – theoriegeleitete Testentwicklung und -evaluation.

Auf der Entwicklung solcher Kompetenztests sollte besonderes Augenmerk liegen, weil Kompetenzmodelle durch die Transparenz von Konstruktionsprinzipien und durch die Operationalisierung in Testaufgaben konkretisiert und

enger an den Unterricht angebunden werden können (Bernholt, Parchmann & Commons, 2009; Klieme, Avenarius et al., 2007). Ein Test wird dabei als Verfahren verstanden, mit dem psychische Eigenschaften oder Merkmale von Personen unter standardisierter Durchführung und Auswertung erfasst werden (Lienert & Raatz, 1998). Im Kontext der Kompetenzmessung wird der Begriff enger gefasst und auf Leistungstests eingegrenzt (Leutner, Hartig & Klieme, 2008). Diese sind aus Skalen zusammengesetzt, die eine Facette des relevanten Konstrukts wie etwa die Teilkompetenz ‚Zweck von Modellen' messen, die wiederum durch Subskalen, z. B. Niveaus innerhalb des ‚Zwecks von Modellen', weiter differenziert werden kann (Leutner et al., 2008; Abb. 6). Die kleinste Beobachtungseinheit ist ein Item, das aus Aufgabenstamm und Antwortformat besteht (Hartig & Jude, 2007; Jonkisz, Moosbrugger & Brandt, 2012; Rost, 2004). Es kann z. B. einem spezifischen Niveau und/oder einer spezifischen Teilkompetenz zugeordnet sein. Dabei können einem Aufgabenstamm, der die Problemstellung beinhaltet, mehrere Items zugeordnet sein (Hartig & Jude, 2007; Leutner et al., 2008).

Für die effiziente Entwicklung eines solchen Tests zu Modellkompetenz mussten zunächst Schritte herausgearbeitet werden, wie der Test und die Items konzipiert werden und welche Aspekte dabei beachtet werden müssen. Terzer et al. (angen.) differenzieren deshalb für die Test- und Itemkonstruktion auf der Grundlage von Kompetenzmodellen sieben Schritte (Abb. 7): 1. Formulierung der theoretischen Fundierung, 2. Testkonzeption, 3. Systematisierung der Itemkonstruktion, 4. Entwicklung einer Konstruktionsanleitung, 5. Itementwicklung, 6. Itemerprobung und –selektion sowie 7. Festlegung des Erhebungsdesigns.

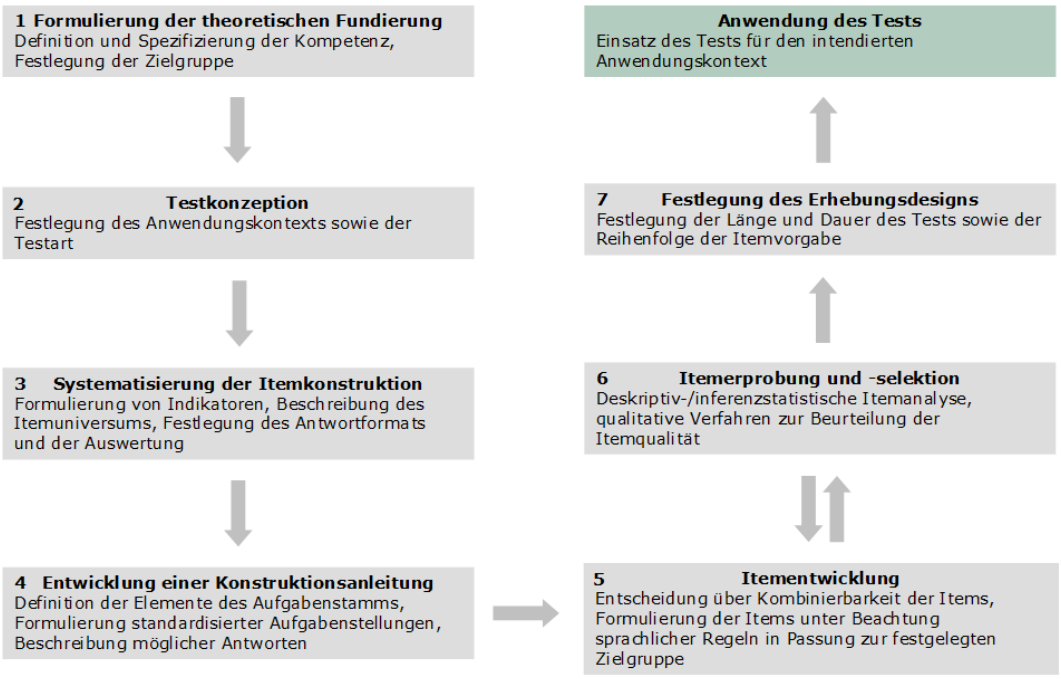


Abb. 7: Schritte der Test- und Itemkonstruktion auf der Grundlage von Kompetenzmodellen (Terzer et al., angen.).

Bereits bei diesen einzelnen Schritten sollten die Gütekriterien Objektivität, Reliabilität und Validität (z. B. Rost, 2004) berücksichtigt werden, da Fehler in diesem Stadium später nur noch durch die Neukonstruktion von Aufgabenmaterial korrigiert werden können (Hartig et al., 2012): "*A large part of test validity must be put into the test at the stage of test construction*" (Borsboom, Mellenbergh & van Heerden, 2004, S. 1067). Schon die Berücksichtigung von einzelnen Schritten der Qualitätssicherung kann als Argument für die valide Interpretierbarkeit der Testwerte herangezogen werden (Haladyna, 1999). Indem man Testziel und -inhalt definiert, eine entsprechende Konstruktionsanleitung formuliert, die Passung zwischen Testziel, Testinhalt und Konstruktionsanleitung prüft und diese Schritte dokumentiert, kann man die Wahrscheinlichkeit erhöhen, dass der entwickelte Test valide interpretierbar ist. Inwiefern dies tatsächlich der Fall ist, muss jedoch laut Osterlind (1998) empirisch geprüft werden, da die Berücksichtigung von Schritten der Testkonstruktion nicht als Evidenz hierfür herangezogen werden kann. Der Fokus liegt im

Folgenden auf der Umsetzung der einzelnen Schritte unter Berücksichtigung der Gütekriterien.

Der erste Schritt der Testentwicklung war die Formulierung einer theoretischen Fundierung (Kapitel 2) sowie entsprechender Strukturmodelle (Kapitel 2.5, H0, H1, H2 in Abb. 5). Sowohl die theoriegeleitete Strukturierung und Graduierung der Kompetenz als auch die Beschreibung von deren Domänenspezifität in einem nomologischen Netzwerk (Köller, 2008a) sind für die Entwicklung eines Instruments zur Erfassung von Modellkompetenz notwendig. Aufbauend darauf wurde systematisch ein Instrument zur Erfassung von Modellkompetenz entwickelt (Kapitel 3.1) und mit Blick darauf evaluiert, ob es das Konstrukt Modellkompetenz angemessen repräsentiert (Kapitel 3.2; Abb. 6). Darüber hinaus muss die Bearbeitung der Items als Indikator von Modellkompetenz interpretiert werden können. Dies ist der Fall, wenn die Items kognitive Prozesse in Bezug auf Modellkompetenz auslösen und diese zur Beantwortung der Items beiträgt (Kapitel 3.3; Abb. 6). Diese Schritte erlauben die abschließende Beurteilung entwickelten Tests als Operationalisierung des Kompetenzmodells (Kapitel 3.4). Im Anschluss daran konnte ein Erhebungsdesign festgelegt und das entwickelte, evaluierte Instrument zur empirischen Beschreibung von Modellkompetenz eingesetzt werden (Kapitel 4).

3.1 Systematische Instrumentenentwicklung

Für die systematische Entwicklung eines Instruments muss mit der Testkonzeption zunächst eine Richtung festgelegt werden (Kapitel 3.1.1). Dies betrifft Ziel und Art des Tests. Die Inhalte des Modellkompetenz-Tests wurden durch das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) definiert (*test content specification*; Osterlind, 1998).

Um die Konstruktion entsprechender Items zu systematisieren, wurden ein Antwortformat festgelegt (hier MC-Items), ein Gegenstandsbereich für die Items sowie jeweils zu erfassende Fähigkeiten definiert und Aufgabenstellungen je Teilkompetenz und Niveau formuliert (Kapitel 3.1.2). Diese Entscheidungen wurden in einer Konstruktionsanleitung standardisiert dokumentiert (Kapitel 3.1.3), so dass systematisch Items entwickelt wurden und somit der

Bezug zwischen Kompetenz und Operationalisierung in einem entsprechenden Testinstrument in jedem Schritt nachvollziehbar war. Der Validitätsaspekt, dass die Beantwortung der Items den Schluss auf ein nicht beobachtbares, latentes Konstrukt – hier Modellkompetenz – ermöglichen soll, wurde somit „durch eine gute theoretische Fundierung, eine daran orientierte Itementwicklung und eine schlüssige Argumentation" (Hartig et al., 2012, S. 150) in eine frühe Phase der Itemkonstruktion eingebunden. Inwiefern die Aufgabenmerkmale und insbesondere die Indikatoren für Teilbereiche der Kompetenz sinnvoll gewählt sind und somit die angesteuerte Kompetenz repräsentieren, wurde durch Expertenurteile evaluiert (vgl. Rost, 2004; Hartig & Jude, 2007; Kapitel 3.1.4).

Auf der Grundlage der evaluierten Konstruktionsanleitung als *test item specification* (Osterlind, 1998) wurden in Orientierung an Leitlinien (z. B. Haladyna, 1999; Rost, 2004) Items entwickelt (Kapitel 3.1.5). Hierfür wurden gezielt Kontexte ausgewählt, Antwortmöglichkeiten aus Schülerantworten generiert und die sprachliche Qualität der Formulierungen geprüft. Die MC-Items wurden in einem nächsten Schritt sukzessive in mehreren Teilstudien empirisch erprobt und für die empirische Beschreibung von Modellkompetenz selektiert (Kapitel 3.1.6). Dabei wurden als Kriterien Schwierigkeit (z. B. Kelava & Moosbrugger, 2012), Trennschärfe (z. B. Kelava & Moosbrugger, 2012) sowie in der IRT-Skalierung *Itemfit* (Wu & Adams, 2007) und ICC (Wu & Adams, 2007) herangezogen.

3.1.1 Testkonzeption

Zur Testkonzeption ist zunächst eine Entscheidung über das Ziel und den Kontext, in dem die relevante Kompetenz erfasst werden soll, notwendig, da sich hieraus Anforderungen für den Test ableiten. Der Forschungsbereich Kompetenzmodellierung ist mit vielfältigen Zielen und damit vielfältigen Anwendungskontexten verbunden (Kapitel 2.1). Für Projekte der Grundlagenforschung zur empirischen Beschreibung von Kompetenzen, zu denen das hier vorgestellte Projekt zählt, steht eher die angemessene Abbildung des Gegenstandsbereichs im Test durch eine große inhaltliche Breite im Vordergrund als eine große Messgenauigkeit auf Individualebene (Klieme, Avenarius et al.,

2007; Koeppen et al., 2008). Die hier beschriebene Operationalisierung des Kompetenzmodells der Modellkompetenz hatte zunächst als Ziel, Modellkompetenz empirisch beschreiben zu können. Darüber hinaus sollte sie im Anschluss auf Klassenebene Aussagen darüber ermöglichen, inwiefern Interventionen erfolgreich sind, und Lehrerinnen und Lehrern Rückmeldungen zum *Outcome* ihres Unterrichts geben können. Mit Blick auf die Bildungsstandards (KMK, 2005), die Aspekte von Modellkompetenz beinhalten, bildeten Schülerinnen und Schüler der Jahrgangsstufen 7 bis 10 von Sekundarschule und Gymnasium die Zielgruppe des Tests.

Instrumente zur Kompetenzmessung sollten Inferenzschlüsse auf die Leistungsfähigkeit in relevanten, domänenspezifischen Situationen zulassen (Koeppen et al., 2008). In Bezug auf Modelle in der Biologie betrifft dies den konkreten Umgang mit Modellen: Wie werden konkrete Modelle in spezifischen Situationen angewendet? Welche Vorstellungen über den Bezug zwischen Modell und Original sowie die Begründung alternativer Modelle können dazu herangezogen werden? Leisner-Bodenthin (2006) sowie Leisner und Mikelskis (2004) unterscheiden domänenübergreifendes und -spezifisches Wissen, die beide durch das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) abgedeckt werden (vgl. Kapitel 2.4.1). Das hier vorgestellte Projekt operationalisiert die domänenspezifische Facette von Modellkompetenz. Der Fokus liegt dabei auf der kognitiven Facette der konkreten Anwendung von Modellen (vgl. Kapitel 2.4.1). Um Modellkompetenz zu erfassen und Modellkompetenz empirisch strukturieren zu können, wurde ein Leistungs-Powertest (z. B. Amelang & Zielinski, 2002; Jonkisz et al., 2012) entwickelt. Kognitive und motivationale Facetten von Kompetenz können ausschließlich getrennt voneinander valide erhoben werden (Köller, 2008a), so dass für motivationale Anteile von Modellkompetenz ein weiteres Instrument entwickelt werden muss.

3.1.2 Systematisierung der Itemkonstruktion

Eine weitere grundlegende Entscheidung im Rahmen der Systematisierung der Itemkonstruktion betrifft das Antwortformat. In zwei weiteren Projekten wurde das Kompetenzmodell durch FC-Items (Krell & Krüger, 2010) und Aufgaben

in offenem Antwortformat (Grünkorn & Krüger, 2012) operationalisiert, im vorliegenden durch MC-Items. Dies ermöglicht in einer gemeinsamen Datenerhebung den Vergleich verschiedener Antwortformate.

MC-Items lassen als zeitökonomisches Testformat den Einsatz einer größeren Anzahl an Items je Niveau und Teilkompetenz zu. Dadurch können die formulierten Niveaustufen über eine Varianzanalyse der Itemschwierigkeiten in Abhängigkeit vom jeweils zugeordneten Niveau und der theoriegeleitet angenommenen Struktur über IRT-Modelle sinnvoll geprüft werden (Kapitel 2.1.2, 4). Dieser Itemtyp weist eine hohe Auswertungsobjektivität auf – problematisch ist jedoch, dass MC-Items häufig einfacher zu beantworten sind als Aufgaben mit einem offenen Format, die das eigenständige Produzieren einer Lösung erfordern (Jonkisz et al., 2012; Martinez, 1999), so dass die Validität von MC-Items zu prüfen ist (Hartig & Jude, 2007). Für die angestrebte empirische Strukturierung und Beschreibung von Modellkompetenz war z. B. relevant, inwiefern die Items das Kompetenzmodell angemessen operationalisieren. In der Forschung zu Vorstellungen zu Modellen wurden jedoch bereits erfolgreich geschlossene Itemformate eingesetzt, so dass anzunehmen ist, dass diese Inhalte grundsätzlich geschlossen erhoben werden können (vgl. Kapitel 2.3.1).

Die kognitiven Anforderungen von Items werden bei Kompetenzmodellen in Form der Zuordnung von Items zu Kompetenzniveaus eines Niveaumodells und/oder zu Dimensionen eines Strukturmodells definiert (Terzer et al., angen.). Die MC-Items wurden je Teilkompetenz und Niveau entwickelt, während sich die FC-Items sowie die Items in offenem Antwortformat jeweils auf alle drei Niveaus innerhalb einer Teilkompetenz beziehen (Grünkorn & Krüger, 2012, Krell & Krüger, 2010). Entsprechend wurden im vorliegenden Projekt für jedes Niveau jeder Teilkompetenz Indikatoren formuliert und in die Konstruktionsanleitung aufgenommen, in denen sich die individuelle Ausprägung der jeweiligen Kompetenz zeigt (vgl. Mislevy & Haertel, 2006; Anhang 2). Somit wurde der inhaltliche Geltungsbereich eines Items auf jeweils ein Niveau einer Teilkompetenz eingegrenzt, so dass sich Antwortmöglichkeiten formulieren

ließen, die verschiedene Vorstellungen zu Modellen bzw. denkbare Anwendungen eines Modells abbilden.

Der inhaltliche Geltungsbereich betrifft darüber hinaus biologische Inhalte der Items. Sowohl mit Blick auf die potenzielle Validität von MC-Items zu Modellkompetenz (vgl. Kapitel 2.3.1) als auch mit Blick auf den hier zugrunde gelegten Kompetenzbegriff (vgl. Kapitel 2.1.1) sollten domänenspezifische, kontextgebundene Items eingesetzt werden. Diese sollten sich zum einen auf konkrete biologische Inhalte beziehen. Für die Befragung einer breiten Zielgruppe müssen die Items über einen großen Schwierigkeitsbereich streuen und inhaltlich breit gefächert sein, um möglichst viele Merkmalsausprägungen abdecken zu können (Jonkisz et al., 2012), so dass zu verschiedenen biologischen Themen (u. a. Evolution, Humanbiologie, Ökologie, Zoologie, Botanik) Items entwickelt wurden (vgl. Campbell, 2009). Mit Blick auf die Funktionalität des Kompetenzbegriffs (vgl. Kapitel 2.1.1) wurden die Items in möglichst realistische, schülernahe Situationen eingebettet, die im Umgang mit Modellen relevant sind. Solche Items werden darüber hinaus von Schülerinnen und Schülern häufig als interessanter wahrgenommen, was die Testmotivation positiv beeinflusst (Hammann, 2006). Sowohl der fachliche Inhalt als auch die Auswahl der Situationen sind schwierigkeitserzeugend (Prenzel, Häußler, Rost & Senkbeil, 2002; Schecker & Parchmann, 2006).

Für die weitere Systematisierung der Itemkonstruktion, die den systematischen Einfluss nicht-theoriegeleiteter Merkmale minimieren soll, unterscheiden Terzer et al. (angen.) darüber hinaus folgende formale Merkmalsbereiche:

- *formale Merkmale* wie die Textlänge und die Einbindung von Abbildungen (z. B. Jonkisz, Moosbrugger & Brandt, 2012; Rost, 2004),
- die zur Lösung des Items notwendige *Wissensbasis* (z. B. Prenzel et al., 2002) sowie
- den *Aufgabenkontext*, der möglichst alltagsnah in Situationen aus dem Erfahrungsbereich der getesteten Personen eingebettet sein sollte (z. B. Hammann, 2006).

Die MC-Items sollen kein Fachwissen messen, da Modellkompetenz im Bereich der Erkenntnisgewinnung zu verorten ist (vgl. Kapitel 2.4). Entsprechend wurde das Fachwissen, das zur Lösung der Items notwendig ist, im Aufgabenstamm vorgegeben. Um eine Konfundierung mit Lesefähigkeiten zu vermeiden und den Einfluss der Textlänge (Prenzel et al., 2002) sowie des Wortschatzes (Hartig & Klieme, 2006; Rost, 2004) zu minimieren, wurden im Stamm möglichst kurze, einfache Texte verwendet (vgl. Haladyna, 1999).

Diese systematische Beschreibung des „Itemuniversums", das heißt der Items, die relevante Fähigkeiten und ihre interindividuellen Unterschiede erfassen (Wilson, 2005), ermöglicht es, das Testergebnis nicht nur in Bezug auf die im Test enthaltenen Items zu interpretieren. Es legt außerdem fest, auf welche Bereiche der Kompetenz ein Testergebnis Schlüsse zulässt (Rost, 2004). Insgesamt wurde bei der Konzeption der Items darauf geachtet, dass sie für die angestrebte Zielgruppe (Kapitel 2.1.2) geeignet sind, so dass die Varianz der Antworten maximiert wird (Rost, 2004).

3.1.3 Entwicklung einer Konstruktionsanleitung

Die Beschreibung von Itemmerkmalen in einer Konstruktionsanleitung hat das Ziel, einen möglichst großen Anteil der Varianz in den Schülerantworten theoriegeleitet erklären zu können und den Entwicklungsprozess der Items ausgehend von der theoretischen Fundierung nachvollziehbar zu machen. Somit kann bereits bei der Entwicklung von Items systematisch argumentiert werden, warum vom Testverhalten auf die individuelle Ausprägung des Konstrukts geschlossen werden kann (Hartig et al., 2012). Darüber hinaus wird die Entwicklung einer Konstruktionsanleitung empfohlen, wenn mehrere Personen an der Itementwicklung beteiligt sind (Haladyna, 1999; Osterlind, 1998). Drei erfahrene Lehrerinnen und Lehrer ergänzten den Aufgabenpool auf Grundlage ihrer Erfahrung aus dem Schulalltag. Die Einbeziehung mehrerer Entwicklerinnen und Entwickler sollte eine breite Themenvielfalt, Praxisnähe und die Eignung der Items für die Zielgruppe sicherstellen. Die Konstruktionsanleitung steuerte die Itementwicklung gleichzeitig so, dass dabei klare Bezüge zur theoretischen Fundierung (Kapitel 2) gewährleistet waren.

Für eine Konstruktionsanleitung nennen Terzer et al. (angen.) folgende Elemente:

- eine Beschreibung der Informationen, die im *Aufgabenstamm* enthalten sein sollen,
- die Formulierung eines standardisierten *Aufgabenimpulses* für jedes Element der Theorie, zu dem Items entwickelt werden,
- eine Beschreibung der *Antwortmöglichkeiten* bzw. des Erwartungshorizontes und
- die Beschreibung der *Kompetenz*, die mit dem Item erfasst werden soll, sowie ein entsprechender *Indikator*.

Entsprechend wurden je Teilkompetenz und Niveau standardisierte Fragen entwickelt (Tab. 4, Tab. 5), um mit jedem Item für einen Bereich des Kompetenzmodells die gleichen Inhalte abzufragen und eine Variation der Schwierigkeit durch die Formulierung der Frage auszuschließen (vgl. Grube, 2011, Kapitel 2.1.2). Ein Beispiel für eine Itembeschreibung in der Konstruktionsanleitung zeigt Tab. 6, für die vollständigen Itembeschreibungen siehe Anhang 2.

Tab. 4: Operationalisierung der Dimension ‚Kenntnisse über Modelle' in standardisierten Fragen.

Teilkompetenz	Niveau	Standardisierte Frage
Eigenschaften von Modellen	Niveau I	In welcher Eigenschaft stimmt das Modell mit dem Original überein?
	Niveau II	Welche Eigenschaften vom Original sind im Modell vereinfacht?
	Niveau III	Welche Annahme über das Original war die Grundlage für die Entwicklung des Modells?
Alternative Modelle	Niveau I Niveau II Niveau III	Warum gibt es unterschiedliche Modelle zu diesem Original?

Tab. 5: Operationalisierung der Dimension ‚Modellbildung' in standardisierten Fragen.

Teilkompetenz	Niveau	Standardisierte Frage
Zweck von Modellen	Niveau I	Was kann man mit diesem Modell zeigen?
	Niveau II	Welchen Zusammenhang kann man mit diesem Modell erklären?
	Niveau III	Welche Vermutungen über das Original kann man aus diesem Modellversuch ableiten?
Testen von Modellen	Niveau I	Wie kann man prüfen, ob man dieses Modell einsetzen kann?
	Niveau II	Welches dieser Modelle kann für die Fragestellung genutzt werden?
	Niveau III	Wie kann man diese Vermutung testen?
Ändern von Modellen	Niveau I	Was muss man an diesem Modell verändern, damit der Zweck noch erfüllt ist?
	Niveau II	Was muss man nach dieser Entdeckung an diesem Modell verändern?
	Niveau III	[Beschreibung, dass die dem Modell zugrunde liegende Hypothese falsifiziert wurde] Was muss man deshalb am Modell verändern?

Tab. 6: Exemplarische Itembeschreibung zum Zweck von Modellen, Niveau II, aus der Konstruktionsanleitung.

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Welchen Zusammenhang kann man mit [dem Modell] erklären?** Beispiel: Welchen Zusammenhang kann man mit dem (K)Automaten erklären?
Antwortmöglichkeiten	denkbare **bekannte Zusammenhänge im Original**
Kompetenz	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären
Indikator	Die Schülerinnen und Schüler benennen, dass der (K)Automat korrelative Zusammenhänge zwischen Kaubewegung, Speichelfluss und Körpertemperatur erklärt.

Am Beginn der Konstruktionsanleitung wird ein Beispiel-Aufgabenstamm, der „(K)Automat", gezeigt (Anhang 2).

Die Items sollten die Schülerinnen und Schüler in die Rolle von Forscherinnen und Forschern versetzen. Entsprechend mussten alle in den Items verwendeten Modelle das Potenzial zur Generierung von Erkenntnissen haben und durften nicht in erster Linie Vorstellungen vom Modell als Medium aktivieren. Dies gilt häufig für Modelle, die aus der Schule bekannt sind. Die Kontexte der Items mussten fachlich so einfach sein, dass sie für Schülerinnen und Schüler ab der Jahrgangsstufe 7 in einem möglichst kurzen Text erklärbar sind. Das biologische Fachwissen, das Schülerinnen und Schüler ggf. zur Lösung des Items benötigen, wurde im Itemstamm beschrieben, um einen Einfluss dieser Variable zu vermeiden (vgl. Kapitel 3.1.2). Der Itemstamm umfasst auch die Informationen über das gezeigte Modell, die lösungsrelevant sind und nicht von der Schülerinnen und Schülern selbst erschlossen werden sollten.

3.1.4 Validierung der Konstruktionsanleitung

Rating

Die Formulierung von Indikatoren, die konkrete kognitive Anforderungen der Items beschreiben, und ihre Überprüfung durch ein Expertenrating sind gängige Forschungspraxis bei der Testkonstruktion auf der Grundlage eines Kompetenzmodells (Schecker & Parchmann, 2006). Um zu überprüfen, inwiefern die Konstruktionsanleitung eine adäquate Operationalisierung des Kompetenzmodells darstellt, wurde ein nominalskaliertes Rating der Itembeschreibungen durchgeführt. Hierfür wurde ein *„blind*" *panel* genutzt, in dem die Raterinnen und Rater ihre Entscheidungen nicht miteinander diskutieren, um eine gegenseitige Beeinflussung zu vermeiden (Osterlind, 1998). Neun Raterinnen und Rater, die in der empirischen Bildungsforschung tätig sind und nicht an der Entwicklung der Items beteiligt waren, ordneten auf der Grundlage einer Einführung in die Theorie zu Modellkompetenz und dem Kompetenzmodell jeweils zehn Itembeschreibungen (Aufgaben- bzw. Itemstamm, Frage sowie mögliche Antworten, vgl. Anhang 2) einer Teilkompetenz sowie einer Niveaustufe des Kompetenzmodells zu. Da sie die intendierte Zuordnung des Items nicht kannten und sich keine vollständige Zuordnung ergab, wurde ein Einfluss hierdurch vermieden (Osterlind, 1998). Jede Itembeschreibung wurde im hier vorgestellten Projekt von sechs Personen dem Kompetenzmodell zu-

geordnet. Dazu wurden den Itembeschreibungen zufällig Nummern zugewiesen. Sowohl die Verteilung der Itembeschreibungen auf die Testhefte für das Rating als auch ihre Reihenfolge innerhalb der Testhefte erfolgten zufällig.

Im Gegensatz zu komplexen Kompetenzmodellen, bei denen die Itembeschreibungen in mehr als zwei Kategorien eingeordnet werden (wie z B. beim Kompetenzmodell zu physikalischen Kompetenzen im Bereich Wärmelehre; Einhaus & Schecker, 2006), wurde hier jedem Item genau eine Kombination aus Teilkompetenz und Niveau zugeordnet. Deshalb kann die Übereinstimmung zwischen theoretischer und empirischer Zuordnung der Konstruktionsanleitung deskriptiv über die prozentuale Übereinstimmung sowie zufallskorrigiert als Cohens κ beschrieben werden (Wirtz & Caspar, 2002). Die prozentuale Übereinstimmung wurde als $PÜ_{gesamt}$ aggregiert für alle Raterinnen und Rater berechnet. Da hier nicht die Übereinstimmung der Raterinnen und Rater untereinander relevant war, sondern die der einzelnen Raterinnen und Rater mit der theoretischen Grundlage der Itembeschreibungen, wurde Cohens κ als Übereinstimmung mit der Theorie für jede Raterin und jeden Rater berechnet und ein Mittelwert der Übereinstimmungen gebildet. Diese Übereinstimmungsmaße können als Hinweis darauf interpretiert werden, inwiefern die Konstruktionsanleitung das Kompetenzmodell nachvollziehbar operationalisiert und somit eine solide Grundlage für die Entwicklung von MC-Items zur empirischen Beschreibung von Modellkompetenz ist.

Empirische Befunde

Bei dem Rating der Konstruktionsanleitung stimmten die angesteuerte und die zugeordnete Teilkompetenz und Niveaustufe für sechs von fünfzehn Itembeschreibungen bei allen Raterinnen und Ratern überein (Tab. 7, für die Rohdaten siehe Anhang 3). Die Zuordnung der Niveaus stellte sich in den folgenden Bereichen als Schwierigkeit heraus: ‚Eigenschaften von Modellen', Niveau I und II, ‚Zweck von Modellen', Niveau I, ‚Testen von Modellen', Niveau I und II, sowie ‚Ändern von Modellen', Niveau III. Die Raterinnen und Rater grenzten die ‚Eigenschaften von Modellen', Niveau I und II, nicht in allen Fällen in Übereinstimmung mit der Theorie zum ‚Testen von Modellen', Niveau II, ab

und ordneten sie in einigen Fällen den jeweils anderen Bereichen zu. Insgesamt lag die Übereinstimmung zwischen theoretischer und empirischer Zuordnung der Konstruktionsanleitung bei $M_k = 0.79$ ($SD_k = 0.13$).

Tab. 7: Ergebnisse des Ratings der Konstruktionsanleitung ($n_{Raterinnen\ und\ Rater\ je\ Teilkompetenz\ und\ Niveaustufe} = 6$, $N = 9$) – prozentuale Übereinstimmung $PÜ_{gesamt}$ der angesteuerten und zugeordneten Teilkompetenz bzw. Niveaustufe.

Teilkompetenz	Niveau	Übereinstimmung mit angesteuerter Teilkompetenz und Niveaustufe [%]	Übereinstimmung mit angesteuerter Teilkompetenz [%]	Übereinstimmung mit angesteuerter Niveaustufe [%]
Eigenschaften von Modellen	I	66.7	83.3	66.7
	II	66.7	100	66.7
	III	83.3	100	83.3
Alternative Modelle	I	100	100	100
	II	66.7	83.3	83.3
	III	100	100	100
Zweck von Modellen	I	66.7	66.7	100
	II	100	100	100
	III	100	100	100
Testen von Modellen	I	66.7	100	66.7
	II	66.7	83.3	83.3
	III	100	100	100
Ändern von Modellen	I	100	100	100
	II	83.3	100	83.3
	III	66.7	100	66.7

Beurteilung des ersten Operationalisierungsschritts

Die Übereinstimmung zwischen theoretischer und empirischer Zuordnung kann nach Wirtz und Caspar (2002) insgesamt als sehr gut bewertet werden. Entsprechend kann angenommen werden, dass die Konstruktionsanleitung insgesamt das Kompetenzmodell angemessen operationalisiert und eine geeignete theoriebasierte Grundlage für die Testkonstruktion darstellt. Auch die prozentuale Übereinstimmung $PÜ_{gesamt}$ der Zuordnung mit der angesteuerten

Teilkompetenz und Niveaustufe von mindestens 66.7 %, bei sechs von fünfzehn Itembeschreibungen von 100 %, spricht hierfür.

Wenn man sich die Ergebnisse im Detail ansieht, sind jedoch einige Auffälligkeiten zu bemerken. Nicht alle Raterinnen und Rater grenzten die Itembeschreibungen zu ‚Eigenschaften von Modellen', Niveau I und II, sowie ‚Testen von Modellen', Niveau II, klar voneinander ab und ordneten Itembeschreibungen teilweise einem der jeweils anderen Bereiche zu. Das Parallelisieren mit dem Original (‚Testen von Modellen', Niveau II), weist theoretisch einen engen Bezug zur Teilkompetenz ‚Eigenschaften' auf, die auf die Beziehung von Modell und Original abzielt (Upmeier zu Belzen & Krüger, 2010). In den drei Bereichen des Kompetenzmodells, welche die Raterinnen und Rater im Vergleich zur theoretischen Grundlage der Itembeschreibung „verwechselten", geht es darum, Modell und Original miteinander zu vergleichen. Niveau I und II im Bereich ‚Eigenschaften von Modellen' stellen einen deutlich engeren Bezug zwischen Modell und Original her, als dies auf Niveau III der Fall ist. Während Niveau I, als Fähigkeit formuliert, die Parallelen von Modell und Original in den Blick nimmt (‚Modelle sind Kopien von etwas'; Upmeier zu Belzen & Krüger, 2010), spricht Niveau II die Grenzen dieses Vergleichs an (‚Modelle sind idealisierte Repräsentationen von etwas'; Upmeier zu Belzen & Krüger, 2010). Dass beide Niveaus einen gegenüber Niveau III (Modelle sind theoretische Rekonstruktionen von etwas; Upmeier zu Belzen & Krüger, 2010) engen Vergleich zwischen Modell und Original erfordern, erklärt die Zuordnung als jeweils anderes Niveau aus theoretischer Sicht und die im Unterschied zu anderen Itembeschreibungen geringere prozentuale Übereinstimmung zwischen theoretischer und empirischer Zuordnung. Da diese Niveaus auf einem unterschiedlich komplexen Wissenverschaftsverständnis beruhen (vgl. Kapitel 2.3.3), sollten diese Bereiche trotz der empirischen und theoretischen Nähe zunächst getrennt und ihre Beziehung auf der Grundlage von Schülerantworten empirisch untersucht werden.

Der Vergleich von Modell und Original spielte bei einer weiteren abweichenden Zuordnung eine Rolle. Eine Raterin und ein Rater ordneten die Itembeschreibung zum ‚Zweck von Modellen', Niveau I, als ‚Eigenschaften', Niveau I, zu.

Auch diese Zuordnung lässt sich durch einen engen theoretischen Bezug zwischen diesen Bereichen erklären: Die Frage, was ein Modell zeigt (vgl. Tab. 5), berührt die Frage, in welchen Eigenschaften das Modell mit dem Original übereinstimmt (vgl. Tab. 4).

Auch im Bereich ‚Testen von Modellen' traten abweichende Zuordnungen auf. Eine Raterin und ein Rater ordneten hier Niveau I-Itembeschreibungen Niveau II zu. Hier handelt es sich möglicherweise um ein leicht abweichendes Verständnis der theoretischen Grundlage: Die Überprüfung des Modellobjekts (Niveau I) wurde als Frage operationalisiert, wie die Einsetzbarkeit des Modells überprüft werden kann (Tab. 5). Wenn jemand dies so versteht, dass hier bereits ein Bezug zum Original hergestellt wird, und nicht berücksichtigt, dass es hier ausschließlich um das Modellobjekt geht (s. entsprechende Itembeschreibung in Anhang 2), ordnet er die Itembeschreibung als Niveau II zu.

Bei der guten gesamten Übereinstimmung ist demnach festzustellen, dass sich enge theoretische Bezüge zwischen den Teilen des Kompetenzmodells empirisch abbilden. Auch Schwarz et al. (2009) beschreiben, dass im konkreten Umgang mit Modellen immer auf Vorstellungen zu Modellen aus verschiedenen Bereichen zurückgegriffen wird. Die explizite Modellierung dieser Bezüge kann jedoch gegenüber einem globalen Bezug, wie ihn Schwarz et al. (2009) herstellen, konkretere Anknüpfungspunkte für die Förderung bieten. Da aus der Theorie heraus die Nähe der empirisch abweichend zugeordneten Bereiche des Kompetenzmodells begründet werden kann und die Operationalisierung in Itembeschreibungen somit Bezüge zwischen Elementen der theoretischen Fundierung abbildet, spricht dies dafür, dass die Operationalisierung die theoretische Grundlage sinnvoll repräsentiert.

Die Nähe der Bereiche ‚Eigenschaften von Modellen', Niveau I und II, sowie ‚Testen von Modellen', Niveau I, und ‚Eigenschaften von Modellen', Niveau I, sowie ‚Zweck von Modellen', Niveau I, könnte sich empirisch bei der Bearbeitung der Items darin ausdrücken, dass diese Bereiche stärker miteinander zusammenhängen als andere. Hierzu wurden ergänzend zu den Hypothesen, die in der Problemstellung (Kapitel 2.5) formuliert wurden, weitere Hypothe-

sen aufgestellt (**H11**, Abb. 8; **H12**, Abb. 9). Auch diese können durch eine explizite Modellierung dieser Bezüge geprüft werden.

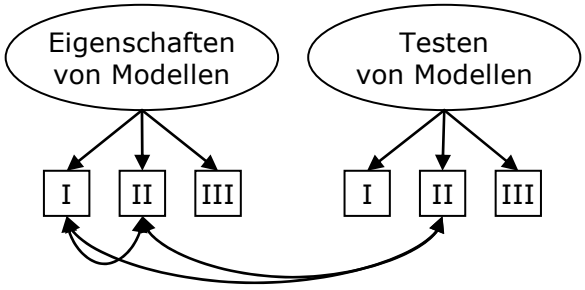


Abb. 8: Strukturmodell zu Hypothese **H11**. Die römischen Ziffern stehen für die Items der jeweiligen Niveaustufen.

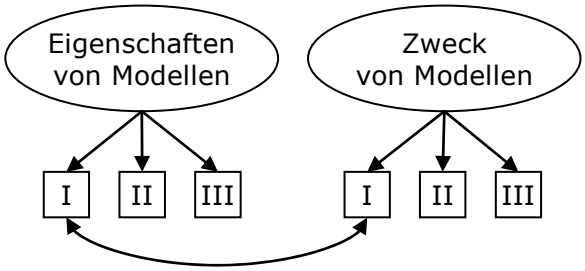


Abb. 9: Strukturmodell zu Hypothese **H12**. Die römischen Ziffern stehen für die Items der jeweiligen Niveaustufen.

3.1.5 Itementwicklung

Die im Expertenrating evaluierte Konstruktionsanleitung diente der Entwicklung von Items für die empirische Beschreibung von Modellkompetenz. Wichtige Schritte, die hierzu thematisiert werden müssen, waren dabei die Auswahl der Kontexte, die Verwendung von offenen Antwortformaten als Grundlage der Itemkonstruktion sowie die sprachliche Formulierung der Items (vgl. Terzer et al., angen.).

Auswahl der Kontexte

Bei der Entwicklung der MC-Items zur Modellkompetenz kamen Inhalte zum Einsatz, die mithilfe eines kurzen Aufgaben- bzw. Itemstamms einfach erklärbar sind (vgl. Kapitel 3.1.1). Aus der theoretischen Fundierung ergeben sich für jede Teilkompetenz und Niveaustufe spezifische Anforderungen an die Modelle, die für entsprechende Items in Frage kommen:

Eigenschaften von Modellen

Da mit den Items zu Niveau I geprüft werden soll, inwiefern Schülerinnen und Schüler Übereinstimmungen zwischen Modell und Original korrekt beschreiben, dürfen Modell und Original sich nicht zu ähnlich sein, damit genug Schüleraussagen für Distraktoren formuliert werden können. Günstig sind demnach eher vereinfachte Modelle, die trotzdem einen klaren Bezug zwischen Modell und Original aufweisen, weil sonst die angesteuerte Niveaustufe verfehlt wird. Dies trifft z. B. auf Drahtmodelle der Wirbelsäule zu. Modelle, die für Niveau II dieser Teilkompetenz genutzt werden, dürfen dagegen aus dem gleichen Grund nicht zu stark vereinfacht sein. Darüber hinaus ist die Verwendung von Strukturmodellen schwierig, da Distraktoren zu solchen Modellen wenig plausibel sind. Hier sollten deshalb eher systemische, wenig vereinfachte Modelle eingesetzt werden wie etwa gegenständliche Modelle von Ökosystemen. Im Bereich von Niveau III müssen Modelle ausgewählt werden, bei denen gut erkennbar ist, dass ihnen eine Hypothese zugrunde liegt. Dies ist in der Regel bei Modellexperimenten der Fall.

Alternative Modelle

Modelle für Niveau I-Items unterscheiden sich nur auf Modellobjektebene in Farbe, Form, Material etc. (Mahr, 2004, 2008a, 2008b; Kapitel 2.2.2). Damit eine ausreichende Anzahl an Schüleraussagen für Distraktoren formuliert werden kann, sollten die Modellobjekte untereinander deshalb möglichst wenige solche Unterschiede aufweisen. Für Items im Bereich von Niveau II müssen Modelle nur der Anforderung genügen, dass sie einen unterschiedlichen inhalt-

lichen Fokus auf ein Original setzen. Da Niveau III die Prüfung verschiedener Hypothesen zu einem Original in den Blick nimmt, ist dies das Auswahlkriterium für Modelle zu entsprechenden Items.

Zweck von Modellen

Für Items, die Niveau I in diesem Bereich erfassen sollen, kann jedes Modell verwendet werden, das ein Original veranschaulicht. Die Menge der Modelle, die hier einsetzbar sind, ist demnach fast unbegrenzt. Im Bereich von Niveau II ist wesentlich, dass die Modelle einen Zusammenhang abbilden. Hier kommen demnach ausschließlich Funktions- und theoretische Modelle zum Einsatz, da sie Zusammenhänge im Original fokussieren. Auch für Niveau III können in erster Linie solche Modelle eingesetzt werden, da sie auf der Grundlage von Daten über das Modell (z. B. Beobachtungen oder in einem Modellexperiment gemessene Daten) eine Vorhersage über das Original ermöglichen müssen.

Testen von Modellen

Bei Modellen, die für den Einsatz in Niveau I-Items geeignet sind, müssen verschiedene strukturelle oder funktionelle Mängel wie Materialfehler oder eine mangelnde Beweglichkeit plausibel sein, von denen jedoch nur einer für eine Fragestellung, unter der das Modell verwendet werden soll, relevant ist. Um Distraktoren für diese Items generieren zu können, sollten hier demnach keine sehr einfachen Modelle verwendet werden. Die Operationalisierung des Vergleichs von Modell und Original (Niveau II) weicht von der der anderen Teilkompetenzen und Niveaustufen ab, da die Kompetenz hier darin besteht, in Abhängigkeit vom Zweck des Modells notwendige Parallelen zwischen Modell und Original zu bestimmen. Deshalb werden vier Modelle gezeigt, von denen eines auf der Grundlage eines Vergleichs zwischen Modell und Original für eine Fragestellung ausgewählt wird. Entsprechend müssen die Modelle sich in der Adäquatheit von Parallelen zum Original unter einer bestimmten Fragestellung unterscheiden. Für Niveau III können ausschließlich Modellexperi-

mente verwendet werden, da Schülerinnen und Schüler hier beurteilen, mit welchem Design eine Hypothese über das Original getestet werden kann.

Ändern von Modellen

Kompetenzen im Bereich ‚Ändern von Modellen', Niveau I, beziehen sich darauf, einen Fehler im Modellobjekt zu beheben. Um mit Items isoliert diese Kompetenzen erfassen zu können und nicht Kompetenzen zum ‚Testen von Modellen', Niveau I, zu konfundieren, muss im Item vorgegeben werden, welchen Fehler das Modellobjekt aufweist. Hier können demnach nur Modelle verwendet werden, bei denen ein Mangel, z. B. im Material oder der Funktionalität, besteht. Ähnliches gilt für das ‚Ändern von Modellen', Niveau II: Der Itemstamm muss hier neue Erkenntnisse über das Original, die eine Änderung des Modells implizieren, beinhalten. Für dieses Niveau bieten sich deshalb historische Modelle oder Modelle der jüngeren Forschung an. Ein Beispiel hierfür sind verschiedene Versionen von DNA-Modellen. Niveau III-Items müssen Modellexperimente beinhalten, die aufgrund einer im Experiment falsifizierten Hypothese zu einer Änderung des Modells führen. Eine wichtige Anforderung ist hier, dass die Hypothese, die sich im Experiment als falsch erweist, dennoch plausibel ist, da sonst das Modellexperiment als Ganzes in Frage zu stellen ist. Aus diesem Grund bieten sich insbesondere Modellexperimente zu Phänomenen an, die nicht direkt beobachtbar sind. Dies gilt etwa für den Einfluss der Körperform auf die Geschwindigkeit eines Lebewesens.

Der gleiche Aufgabenstamm und somit ggf. das gleiche Modell kann prinzipiell für Items in verschiedenen Teilkompetenzen genutzt werden. Mit Blick auf das angestrebte Testheftdesign für die empirische Strukturierung und Beschreibung von Modellkompetenz (Kapitel 4.1) wurde jedoch darauf geachtet, dass keine logischen Abhängigkeiten zwischen den Items bestehen (vgl. Haladyna, 1999) und somit alle Items frei miteinander kombinierbar sind.

Itemkonstruktion ausgehend von offenen Antwortformaten

Bei der Konstruktion MC-Items besteht eine Schwierigkeit in der Formulierung geeigneter Distraktoren, d. h. falscher Antwortalternativen (Jonkisz et al.,

2012). Deshalb empfiehlt Wilson (2005) aus ökonomischen Gründen, geschlossene Items ausgehend von offenen Formaten zu konstruieren und dabei sukzessive Interviews sowie Items in offenem Format ohne bzw. mit Kodierleitfaden zu nutzen. Dieses Vorgehen ist mittelfristig effizienter als die direkte Entwicklung von geschlossenen Formaten ohne Antworten der relevanten Zielgruppe als Grundlage (Wilson, 2005) und wurde für Kompetenzitems bereits von Schmiemann (2010) genutzt. Das hier beschriebene Projekt griff auf Ergebnisse der Studien von Trier und Upmeier zu Belzen (2009; Interviews) sowie Grünkorn und Krüger (2012; Items in offenem Antwortformat, über den Verlauf des Projekts Entwicklung eines Kodierleitfadens) zurück. Entscheidend ist hierbei die Qualität der Distraktoren: "*It does not matter how many distractors one produces for any given MC item but it does matter that each distractor performs as intended*" (Haladyna, 1999, S. 90; vgl. auch Osterlind, 1998). Um Verzerrungen durch mangelhafte Distraktoren zu vermeiden, sollten sie

- disjunkt sein, das heißt sich gegenseitig ausschließen,
- dem Attraktor in oberflächlichen Merkmalen wie Wortzahl und sprachlicher Struktur ähneln und
- typische Fehler repräsentieren, die Personen mit einer niedrigen Kompetenz machen (Haladyna, 1999; Kubiszyn & Borich, 2006; Osterlind, 1998).

Die inhaltliche sowie sprachliche Gestaltung plausibler Distraktoren wird erleichtert, wenn für die Antwortalternativen Schülerformulierungen genutzt werden. Auf diese Weise können außerdem typische Fehler in die Distraktoren einbezogen werden. Dies verringert die Wahrscheinlichkeit, dass die getesteten Personen raten, so dass deren Kompetenz genauer gemessen wird, und vergrößert die diagnostische Information, da die Distraktoren nicht nur falsch sind, sondern mit Blick auf die Kompetenzentwicklung interpretiert werden können. Die Verwendung von Schülerantworten für die Entwicklung von Antwortmöglichkeiten erhöht somit sowohl die Reliabilität als auch die Validität der Aufgaben (Wilson, 2005). Indem in einem ersten Schritt je eine Klasse der Extremgruppen hinsichtlich der erwarteten Leistung (Jahrgangsstufe 7, Realschule, und Jahrgangsstufe 10, Gymnasium, jeweils eine Klasse) die Aufgaben

in einer Version mit offenem Antwortformat bearbeiteten, sollte die Formulierung der Antwortmöglichkeiten möglichst das gesamte Spektrum der Zielgruppe einbeziehen. Um zu vermeiden, dass die Schülerinnen und Schüler sich in ihren Antworten stark auf das jeweilige Fachwissen beziehen oder ausschließlich richtige Antworten formulieren, wurden die Impulse im Vergleich zu den MC-Items für die einzelnen Teilkompetenzen und Niveaus umformuliert, z. B. für Testen von Modellen, Niveau I, „Beschreibe, welche Schwierigkeiten es dabei mit dem Modell geben könnte" statt „Wie kann man prüfen, ob man das Modell einsetzen kann?".

MC-Items werden üblicherweise mit Blick auf ihre Reliabilität mit drei bis fünf Antwortmöglichkeiten konstruiert (Osterlind, 1998). Auf der Basis der Schülerantworten wurden deshalb sprachlich und inhaltlich geeignete Attraktoren und jeweils drei Distraktoren formuliert (Tab. 8). Expertinnen und Experten aus der Didaktik der Biologie, die über Erfahrung im Bereich der Aufgabenentwicklung verfügen, diskutierten die fachliche Richtigkeit der Items und die Eignung der ausgewählten biologischen Kontexte für Schülerinnen und Schüler in den Jahrgangsstufen 7 bis 10, so dass Items ggf. geändert werden konnten. In Abgrenzung zu den AAAS-Items, die zwei Vorstellungen zu vier Antwortmöglichkeiten kombinieren (AAAS, o. J.; Kapitel 2.3.1) wurden dabei vier Inhalte für die Anwortmöglichkeiten ausgewählt, die möglichst optimal der theoretischen Grundlage entsprechen. Außerdem wurde die Position des Attraktors variiert, um die Beantwortung der Items nicht durch sie zu beeinflussen (vgl. Haladyna, 1999). Auf diese Weise wurde der Einfluss von Fähigkeiten im Bereich des logischen Schließens verringert, damit die Items nicht nach einer Strategie beantwortet werden, die sich z. B. auf Oberflächenmerkmale der Items statt auf die Modellkompetenz der Schülerinnen und Schüler bezieht. Aus dem gleichen Grund wurden Formulierungen wie „immer", „nie" usw. vermieden (Osterlind, 1998).

Tab. 8: Beispiele für Antworten von Schülerinnen und Schülern und die darauf aufbauende Formulierung von Antwortmöglichkeiten (Terzer et al., angen.).

Item im offenen Antwortformat: **Nenne Zusammenhänge, die man mit diesem Speiseröhren-Modell erklären kann.** Schülerantworten	Multiple-Choice Item: **Welchen Zusammenhang kann man mit diesem Speiseröhren-Modell erklären?** Antwortmöglichkeiten
„Gewicht der Nahrung und Elastizität der Speiseröhre" „Die Fähigkeit, sich zu dehnen und vielleicht auch zusammenzuziehen, erklärt, wie große, schwere Mengen Speisebrei geschluckt werden können."	Den Zusammenhang zwischen dem Gewicht der Nahrung und der Verformbarkeit der Speiseröhre
„Die Speiseröhre ist eine enge Öffnung, bei der die Nahrung nach dem Schlucken entlang rutscht."	Den Zusammenhang zwischen dem Schlucken der Nahrung und der Öffnung der Speiseröhre

Sprachliche Qualität der Items

Jonkisz et al. (2012, S. 64) betonen, dass „die Klarheit des sprachlichen Ausdrucks (…) bei der Itemformulierung oberste Priorität" hat, um Fehlinterpretationen und Motivationseinbußen zu vermeiden. Eine klare Aufgabenstellung ist deshalb von zentraler Bedeutung (Lienert & Raatz, 1998; Neuhaus & Braun 2007). Die MC-Items wurden auf der Grundlage von Empfehlungen für die sprachliche Gestaltung von Aufgaben in geschlossenem Antwortformat formuliert (z. B. Jonkisz et al., 2012; Lienert & Raatz, 1998; Neuhaus & Braun, 2007). Da davon auszugehen war, dass Schülerinnen und Schüler mit Migrationshintergrund an den Befragungen teilnehmen, prüfte eine Expertin aus dem Bereich Deutsch als Zweitsprache nach Rösch (2003) die Items auf ihre sprachliche Qualität sowie auf die Verständlichkeit für die Zielgruppe und optimierte sie, wenn notwendig (vgl. Haladyna, 1999). Dies diente außerdem dazu, einen systematischen Einfluss der sprachlichen Komplexität so gering wie möglich zu halten (Hartig & Klieme, 2006; Rost, 2004; vgl. Kapitel 3.1.1).

3.1.6 Itemerprobung und -selektion

Die entwickelten Items wurden empirisch erprobt und für die empirische Beschreibung von Modellkompetenz selektiert. Je Teilkompetenz und Niveau des Kompetenzmodells wurden drei psychometrisch adäquate, valide MC-Items benötigt (insgesamt 45 Items).

Empirische Erprobung der Items

Insgesamt wurden 191 MC-Items entwickelt und sukzessive in mehreren Teilstudien pilotiert (n = 173 bis n = 397, N = 1229; Abb. 10), in denen jedes Item von n = 56 bis 102 Schülerinnen und Schülern beantwortet wurde. Diese Stichproben setzten sich aus Extremgruppen der Zielgruppe hinsichtlich der erwarteten Leistung zusammen (d. h. Jahrgangsstufe 7, Realschule, und Jahrgangsstufe 10, Gymnasium).

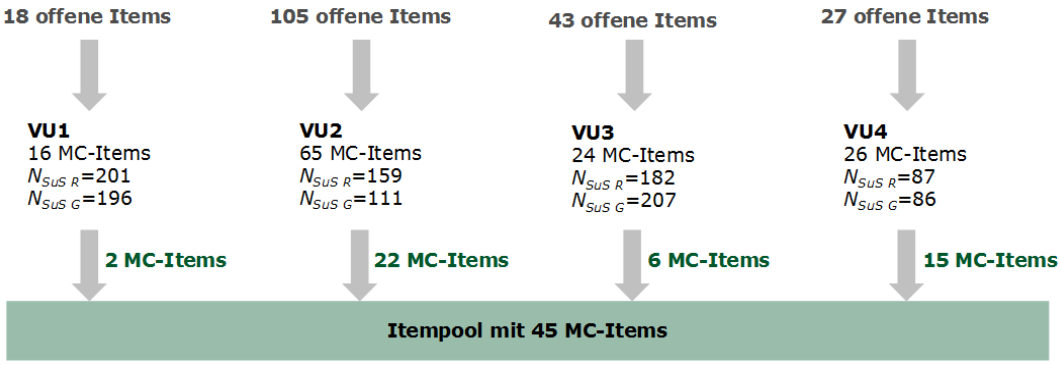


Abb. 10: Überblick über die Pilotierungsteilstudien im Rahmen der Testkonstruktion. SuS = Schülerinnen und Schüler, R = Realschule, G = Gymnasium, VU = Voruntersuchung.

Im Rahmen dieser Itemerprobung wurden die Schülerinnen und Schüler gebeten, Verständnisschwierigkeiten in den Testheften zu markieren und zu kommentieren, so dass die Verständlichkeit durch eine Überarbeitung der Items verbessert werden konnte (vgl. Osterlind, 1998). Neben den Verständnisschwierigkeiten wurden Rückfragen von Schülerinnen und Schülern protokolliert. Dies ermöglichte die Erstellung einer Anleitung zum Einsatz des Frage-

bogens in den Klassen, so dass die Durchführung mit Blick auf Instruktionen und Informationen für die Schülerinnen und Schüler standardisiert wurde, um eine größtmögliche Objektivität der Erhebungssituation gewährleisten zu können (Jonkisz et al., 2012; Neuhaus & Braun, 2007). Darüber hinaus wurden zur Zusammenstellung der Testhefte für die empirische Beschreibung von Modellkompetenz (Kapitel 4.1) Informationen über die Bearbeitungszeit gesammelt, die für die Items benötigt wird (vgl. Osterlind, 1998).

Bei der Evaluation der Items stellte sich heraus, dass alle Items mit physikalischen oder chemischen Bezügen inhaltlich zu anspruchsvoll waren. Solche Kontexte wurden deshalb für die Itementwicklung ausgeschlossen. Für die psychometrische Beurteilung der Items wurden als klassischer Kennwert die Trennschärfe (Übereinstimmung der Differenzierung eines Items mit der Differenzierung der übrigen Items des Tests; z. B. Amelang & Zielinski, 2002; Kelava & Moosbrugger, 2012) berechnet. Darüber hinaus wurden ergänzend drei IRT-basierte Merkmale betrachtet, da die IRT zur empirischen Beschreibung von Modellkompetenz herangezogen wurde:

- Schwierigkeit (Rost, 2004),
- ICC (Wu & Adams, 2007) und
- *Itemfit* (Wu & Adams, 2007).

In Kapitel 4.2.1 wird im Detail beschrieben, inwiefern die IRT die Grundlage der Analysen im hier vorgestellten Projekt bildet. Hier werden deshalb nur die Merkmale beschrieben, die zum Verständnis der Itemselektion notwendig sind. Die ICC beschreibt die Lösungswahrscheinlichkeit eines Items für unterschiedliche Personenfähigkeiten. Sie sollte mit steigender Personenfähigkeit bei MC-Items monoton ansteigen und möglichst nah an der theoretisch vorhergesagten liegen (Wu & Adams, 2007). Inwieweit ein Item zum verwendeten Messmodell, hier ein- bzw. multidimensionale IRT-Modelle, passt, kann außerdem über den *Itemfit* geprüft werden. Dafür werden die vorhergesagten Lösungshäufigkeiten mit den beobachteten verglichen. Der *weighted mean square* (wMNSQ) mit einem Erwartungswert von 1 gibt an, inwiefern die Steigung, d. h. die Trennschärfe eines Items, vom Messmodell vorhergesagt wird. Werte über 1 weisen auf eine flachere ICC und somit eine geringere Trenn-

schärfe des Items als vorhergesagt hin, Werte unter 1 auf eine steilere und damit eine größere Trennschärfe – damit sollten insbesondere Items mit einem wMNSQ unter 1 im Itempool belassen werden (Wu & Adams, 2007). Grundsätzlich sind Werte, die in einem Bereich von $0.75 < wMNSQ < 1.30$ bzw. 1.33 liegen, als akzeptabel einzuordnen (Bond & Fox, 2007; Wilson, 2005). Entsprechende t-Werte im Bereich -2 bzw. $-1.96 < t < +2$ bzw. $+1.96$ (Bond & Fox, 2007; Rost, 2004) zeigen an, dass die Abweichung vom Erwartungswert nicht signifikant ist (Wilson, 2005; Wu & Adams, 2007).

Analog zu den Strukturmodellen, die zur empirischen Strukturierung von Modellkompetenz verglichen werden sollen (Kapitel 2.5, 4.2.3), wurden *Itemfit* und ICC in einer ein-, zwei- sowie fünfdimensionalen IRT-Skalierung mit ConQuest (Version 2.0; Wu, Adams & Wilson, 2007) geprüft. Ein Überblick über die psychometrische Qualität der Items, die für die empirische Strukturierung und Beschreibung von Modellkompetenz selektiert wurden, ist in Anhang 4 zu finden. Bei der Berechnung der Trennschärfe ist zu beachten, dass die Bildung eines Testwerts voraussetzt, dass alle Items dasselbe Merkmal messen (Kelava & Moosbrugger, 2012). Dies ist bei mehrdimensionalen Tests wie dem hier vorliegenden nicht gegeben. Aus diesem Grund werden für die Berechnung der Trennschärfe die Items je Teilkompetenz mit ConQuest eindimensional skaliert, so dass die Trennschärfe im Vergleich zu den anderen Items der entsprechenden Teilkompetenz angegeben werden kann.

Für die Einschätzung der Reliabilität konnte nicht klassisch das Verhältnis von Fehlervarianz und „wahrer" Varianz in den individuellen Messwerten herangezogen werden, da die Analysen nicht auf die Individualebene abzielten (vgl. Kapitel 3.1.1). Stattdessen wurde die nach dem verwendeten Messmodell erwartete „wahre" Varianz (d. h. die der a posteriori-Verteilung, Kapitel 4.2.1) zur nach dem Modell empirisch geschätzten Varianz der Personenfähigkeiten in Beziehung gesetzt. Dieser Wert wird im Rahmen der IRT-Analysen als sog. EAP/PV-Reliabilität ausgegeben (Rost, 2004). Auch hierfür wurden eine ein-, zwei- sowie fünfdimensionale Skalierung durchgeführt. Tab. 9 zeigt die EAP/PV-Reliabilität und Varianz für diese Skalierungen in ConQuest. Abb. 11

stellt die Personenfähigkeiten und Itemschwierigkeiten auf einer gemeinsamen Skala als Wright Map dar.

Tab. 9: EAP/PV-Reliabilität und Varianz für verschiedene Skalierungen in ConQuest.

	Dimensionen des jeweiligen Messmodells							
	1	2		5				
		KM	MB	E	A	Z	T	Ä
Reliabilität	.455	.336	.412	.308	.272	.297	.324	.331
Varianz	.466	.541	.582	.579	.579	.765	.612	.797

Da der vollständige selektierte Itempool (N_{Items} = 45) ausschließlich in der Datenerhebung für die empirische Beschreibung von Modellkompetenz eingesetzt wurde, bildet diese Erhebung die Datengrundlage für die hier gezeigten Werte. Jedes Testheft enthielt neun Items. KM = Kenntnisse über Modelle; MB = Modellbildung; E = Eigenschaften von Modellen; A = Alternative Modelle; Z = Zweck von Modellen; T = Testen von Modellen; Ä = Ändern von Modellen.

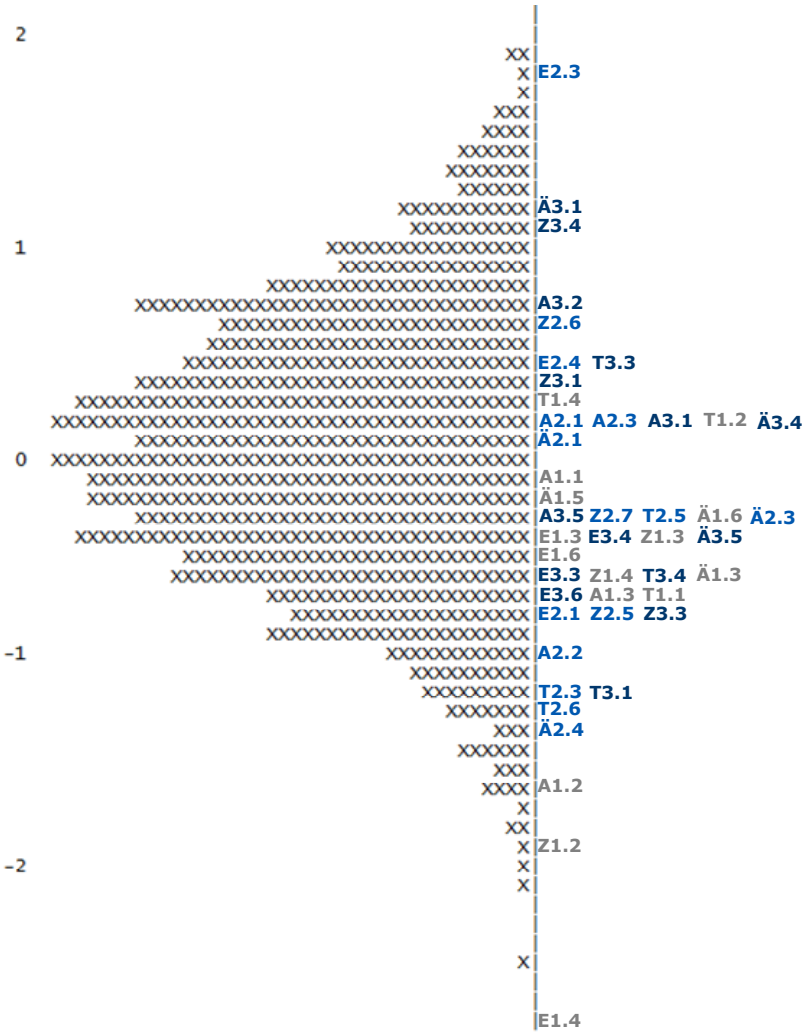


Abb. 11: Wright Map mit Personenfähigkeiten (linke Seite) und Itemschwierigkeiten (rechte Seite) auf einer Logit-Skala[7].

---

[7] Da der vollständige selektierte Itempool ausschließlich in der Datenerhebung für die empirische Strukturierung und Erfassung von Modellkompetenz eingesetzt wurde, bildet diese Erhebung ($N_{Schülerinnen\ und\ Schüler}$ = 1136) die Datengrundlage für die hier gezeigte Wright Map. Die Kürzel für die Items entsprechen den Bezeichnungen in Anhang 6.

Itemselektion

Die MC-Items wurden mit Blick auf die Gesamtheit der oben genannten Kennwerte für die empirische Beschreibung von Modellkompetenz auf drei Items je Teilkompetenz und Niveau (insgesamt 45 Items) reduziert. Häufig war ein Item z. B. statistisch nicht nur zu schwierig, sondern hatte außerdem einen oder zwei sehr attraktive Distraktoren sowie ein oder zwei sehr unattraktive und war nicht zufriedenstellend trennscharf. Gleiches gilt für die IRT-geleiteten Kriterien. Die Selektion der Items erfolgte anhand des Gesamtbildes aller Kriterien, um Ursachen für die darin abgebildeten Probleme zu finden und sie evtl. durch eine Überarbeitung des Items zu beheben (Terzer et al., angen.). Tab. 10 gibt einen Überblick darüber, welche Kennwerte der resultierende Itempool aufweist (für Kennwerte der einzelnen Items siehe Anhang 4).

Tab. 10: Kennwerte des selektierten Itempools zu Modellkompetenz.

Kennwert		Bereich
Schwierigkeit		-2.535 – 1.762
Trennschärfe		0.28 – 0.65
Itemfit	wMNSQ	0.89 – 1.18
	T	-1.2 – 1.7

Der *Itemfit* (wMNSQ, T-Wert) war als Kriterium aufgrund der geringen Varianz nur begrenzt aussagekräftig. Da aufgrund des *Itemfits* jedoch keine Items ausgeschlossen worden wären, die ansonsten gute Kriterien aufwiesen, spielte dieses Kriterium für die Selektion eine untergeordnete Rolle. Stattdessen wurde die Passung zum Mesmodell in einer ein-, zwei- und fünfdimensionalen Skalierung grafisch über die ICC beurteilt. Abb. 12 zeigt exemplarisch die ICC eines Items, das gut zum verwendeten Messmodell passt, Abb. 13 die ICC eines weniger gut geeigneten Items.

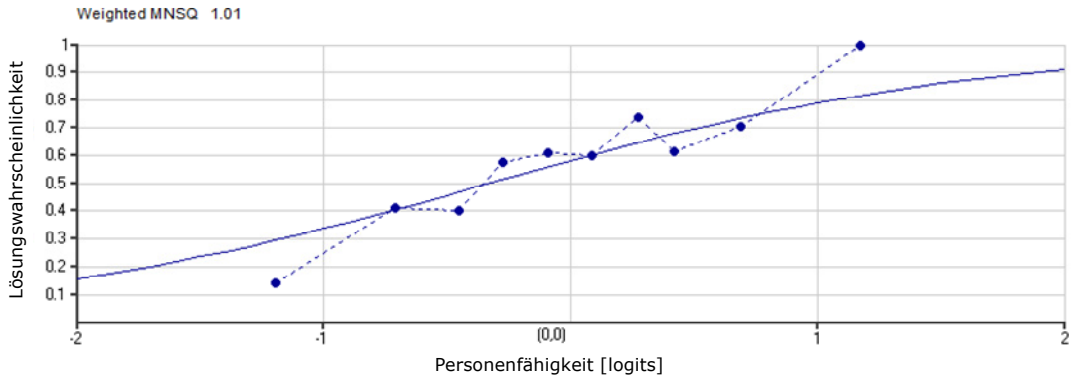


Abb. 12: ICC für das Item Ä2.3 in der eindimensionalen Skalierung. Die empirische ICC ist gestrichelt eingezeichnet, die theoretische als durchgehende Linie.

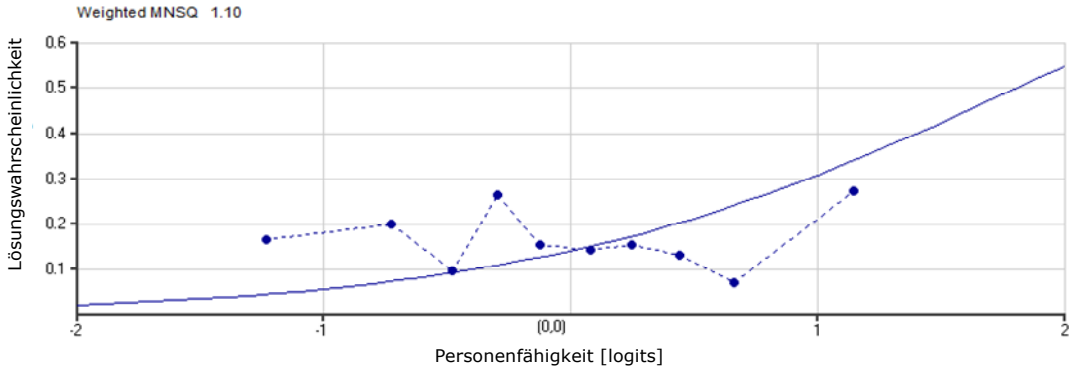


Abb. 13: ICC für das Item E2.3 in der eindimensionalen Skalierung. Die empirische ICC ist gestrichelt eingezeichnet, die theoretische als durchgehende Linie.

Beurteilung des entwickelten Tests

Der entwickelte Test ist mit Blick auf die Verteilung der Personenfähigkeiten als eher leicht einzustufen. Mehr Items an den Rändern der Verteilung wären wünschenswert. Insgesamt deckt der Test das Spektrum der Personenfähigkeiten aber sinnvoll ab, so dass er geeignet ist, um Personenfähigkeiten zu erfassen. Auch wenn die Items sich nicht vollständig entsprechend der Ni-

veaus auf der Skala anordnen, sind Schwerpunkte der Niveaus anhand der Itemschwierigkeiten erkennbar.

Typischerweise nehmen Personenfähigkeiten und Itemschwierigkeiten Werte im Bereich von -3 bis + 3 an (Moosbrugger, 2012). Es stellt sich deshalb die Frage, wie im vorliegenden Projekt die geringere Varianz zu erklären ist. Eine mögliche Erklärung hierfür könnte sein, dass der Test die Modellkompetenz von Schülerinnen und Schülern nicht angemessen erfasst. Da dies bereits im Verlauf der Itementwicklung hinterfragt und in einem Experten-Rating zur Konstruktionsanleitung in einem ersten Ansatz geprüft wurde (Kapitel 3.1.4), ist dies zunächst nicht anzunehmen. Im Rahmen der Testentwicklung wurde im Anschluss an die Itemselektion außerdem geprüft, inwiefern die einzelnen entwickelten Items das Kompetenzmodell angemessen repräsentieren (Kapitel 3.2) und inwiefern die Itembearbeitung Schlüsse auf die Modellkompetenz zulassen und sich die kognitiven Prozesse während der Itembearbeitung auf Modellkompetenz beziehen (Kapitel 3.3). Dies liefert nicht nur Hinweise über die Validität der Items, sondern auch Hinweise, inwiefern die geringe Varianz auf den Test selbst zurückzuführen ist. Ein weiterer denkbarer Grund für die geringe Varianz ist, dass es das postulierte Konstrukt Modellkompetenz empirisch nicht gibt. Da das Kompetenzmodell, auf dem der entwickelte Test basiert, sich wiederum auf empirische Untersuchungen stützt (z. B. Grosslight et al., 1991, Crawford & Cullin, 2005, Justi & Gilbert, 2003; vgl. Kapitel 2.3, 2.4.1), ist auch diese Begründung eher unwahrscheinlich. Darüber hinaus ist bereits an der *Wright Map* der selektierten Items (Abb. 11) zu erkennen, dass die Niveaus vermutlich einen Beitrag zur Erklärung der Itemschwierigkeit leisten. Auch dies spricht für die empirische Existenz von Modellkompetenz. Als dritte mögliche Erklärung bleibt deshalb deren mangelnde Implementierung in den Biologieunterricht. Der empirische Befund, dass Lehrerinnen und Lehrer über Vorstellungen zu Modellen verfügen, die mit denen von Schülerinnen und Schülern vergleichbar sind (Van Driel & Verloop, 1999), stützt diese Vermutung. Es könnte demnach sein, dass Schülerinnen und Schüler sich in einem geringen Umfang in ihrer Modellkompetenz unterscheiden, weil es wenige entsprechende Lerngelegenheiten hierzu gibt.

Bei einer geringen Varianz ist eine hohe Reliabilität nicht zu erwarten, da diese mit der Varianz in Zusammenhang steht und eine entsprechende Testlänge unter Verwendung homogener Items erfordert (z. B. Amelang & Zielinski, 2002). Auch in anderen Studien zur Erfassung naturwissenschaftlicher Kompetenzen werden geringe Reliabilitäten aufgrund geringer Itemanzahlen je Person berichtet (z. B. Schmiemann, 2010; Senkbeil, Rost, Carstensen & Walter, et al., 2005; Wellnitz, 2012). Bereits bei dieser Itemanzahl waren jedoch bei den Schülerinnen und Schülern Schwierigkeiten mit der Konzentration und Motivation zu beobachten, so dass nicht davon ausgegangen werden kann, dass ein deutlich längerer Test zuverlässig konstruktgemäß bearbeitet würde (*Optimizing-Satisficing*-Problem; Jonkisz et al., 2012). Eine höhere Fehlervarianz erschwert den inferenzstatistischen Nachweis von Gruppenunterschieden, dennoch werden die Gruppenmittelwerte auch bei individuell stärker messfehlerbehafteten Testwerten korrekt geschätzt (Schermelleh-Engel & Werner, 2012). Da der Test nicht auf eine Individualdiagnose, sondern auf die Klassenebene abzielt, ist seine Verwendung für die empirische Beschreibung von Modellkompetenz trotz der geringen Reliabilität vertretbar, wenn weder der Test selbst noch die empirische Existenz des Konstrukts zu der geringen Varianz führen.

3.2 Repräsentation des Kompetenzmodells durch die Items

Um die selektierten MC-Items zur empirischen Beschreibung von Modellkompetenz einsetzen zu können, müssen sie dieses Konstrukt angemessen repräsentieren (Hartig et al., 2012; vgl. auch Terzer et al., angen.). Es ist demnach notwendig, die einzelnen Items dem Kompetenzmodell zuordnen zu können. Hierzu wurde ein Experten-Rating herangezogen (Kapitel 3.2.1) und untersucht, inwiefern die empirische Zuordnung der Items durch die Expertinnen und Experten mit der theoretischen Zuordnung der Items übereinstimmt (Kapitel 3.2.2). Auf diese Weise wurde geprüft, ob ein tragfähiger Theoriebezug zwischen den einzelnen Items und dem Kompetenzmodell hergestellt werden kann und die einzelnen Items das Kompetenzmodell angemessen repräsentieren (Kapitel 3.2.3).

3.2.1 Rating

Analog zur theoriegeleiteten Überprüfung der Konstruktionsanleitung als Operationalisierung des Kompetenzmodells (Kapitel 3.1.4) wurde untersucht, inwiefern auch die einzelnen Items der theoretischen Grundlage in einem Rating zugeordnet werden können und sie somit adäquat abbilden. Auf der Grundlage einer Einführung in die Theorie zu Modellkompetenz und dem Kompetenzmodell ordneten Raterinnen und Rater aus der empirischen Bildungsforschung (N = 9; $n_{je\ Item}$ = 2), die nicht an der Entwicklung der Items beteiligt waren, in einem *„blind"* *panel* (Osterlind, 1998) jedes Item einer Teilkompetenz sowie einer Niveaustufe des Kompetenzmodells zu. Jede Person erhielt dafür zehn zufällig ausgewählte Items.

Analog zum Rating der Itembeschreibungen (Kapitel 3.1.4) wurde für die einzelnen Items die Übereinstimmung zwischen theoretischer und empirischer Zuordnung deskriptiv über die prozentuale Übereinstimmung $PÜ_{gesamt}$ je Teilkompetenz und Niveau aggregiert für alle Raterinnen und Rater sowie zufallskorrigiert als Cohens κ (mittlere Übereinstimmung zwischen den einzelnen Raterinnen und Ratern und der Theorie) beschrieben (Wirtz & Caspar, 2002; vgl. Kapitel 3.1.4). Die Übereinstimmungsmaße werden als Hinweis darauf interpretiert, inwiefern die Items das Kompetenzmodell nachvollziehbar operationalisieren und somit für die empirische Strukturierung und Beschreibung von Modellkompetenz geeignet sind.

3.2.2 Ergebnisse

Insgesamt ordneten die Raterinnen und Rater die Items zuverlässiger den jeweiligen Teilkompetenzen als den Niveaus zu (Tab. 11, für die Rohdaten siehe Anhang 5). Die Items zum ‚Zweck von Modellen', Niveau I, sowie ‚Ändern von Modellen', Niveau III, wurden von allen theoriegemäß zugeordnet. Eine besonders geringe Übereinstimmung zwischen der empirischen und theoretischen Zuordnung ergab sich bei ‚Eigenschaften von Modellen' sowie ‚Alternativen Modellen', jeweils Niveau I. In beiden Fällen war die Zuordnung zum angesteuerten Niveau problematisch, während zuverlässig zur angesteuerten Teilkompetenz zugeordnet wurde. Die Hälfte der Raterinnen und Rater stufte

die Items zu ‚Eigenschaften von Modellen', Niveau I, als Niveau II ein, eine Raterin als ‚Testen', Niveau II. 66.7 % der Raterinnen und Rater ordneten ‚Alternative Modelle', Niveau I, als Niveau II ein. Umgekehrt wurden von 33.3 % der Raterinnen und Rater Items zu ‚Alternativen Modellen', Niveau II, als Niveau I eingeordnet. Insgesamt lag die Übereinstimmung zwischen theoretischer und empirischer Zuordnung der Items bei $M_{\kappa} = 0.67$ ($SD_{\kappa} = 0.21$).

Tab. 11: Ergebnisse des Ratings der einzelnen Items der Dimension Kenntnisse über Modelle ($n_{Raterinnen\ und\ Rater\ je\ Aufgabe} = 2$, $N = 9$) – prozentuale Übereinstimmung $PÜ_{gesamt}$ der angesteuerten und zugeordneten Teilkompetenz bzw. Niveaustufe.

Teilkompetenz	Niveau	Übereinstimmung mit angesteuerter Teilkompetenz und Niveaustufe [%]	Übereinstimmung mit angesteuerter Teilkompetenz [%]	Übereinstimmung mit angesteuerter Niveaustufe [%]
Eigenschaften von Modellen	I	33.3	83.3	33.3
	II	50.0	66.7	66.7
	III	50.0	50.0	83.3
Alternative Modelle	I	33.3	100	33.3
	II	66.7	100	66.7
	III	66.7	83.3	66.7
Zweck von Modellen	I	100	100	100
	II	83.3	100	83.3
	III	50.0	100	50.0
Testen von Modellen	I	83.3	83.3	83.3
	II	83.3	83.3	83.3
	III	83.3	83.3	100
Ändern von Modellen	I	83.3	100	83.3
	II	83.3	100	83.3
	III	100	100	100

3.2.3 Diskussion

Die Übereinstimmung zwischen theoretischer und empirischer Zuordnung war im Vergleich zum Rating der Konstruktionsanleitung der einzelnen Items geringer, kann nach Wirtz und Caspar (2002) aber als gut angesehen werden. Der Bezug zwischen der theoretischen Grundlage und einzelnen Items war demnach schwieriger herzustellen als der zwischen theoretischer Grundlage und Itembeschreibungen. Eine mögliche Erklärung ist, dass die konkreten Elemente des Aufgabenstamms dazu führen, dass die Raterinnen und Rater bei der Zuordnung des Items weniger auf die Fragestellung sowie mögliche Antworten fokussieren und stattdessen die Abbildungen und Merkmale des Kontexts wie die Präsentation eines Modellexperiments leitend für die Zuordnung sind. Dies spricht jedoch, insbesondere in Anbetracht der immer noch guten Übereinstimmung, zunächst nicht dagegen, dass die Items die theoretische Grundlage adäquat repräsentieren.

Abweichungen in den Zuordnungen der Items kamen, wie beim Rating der Konstruktionsanleitung (Kapitel 3.1.4), vor allem bei den Niveaus innerhalb einer Teilkompetenz vor und weniger zwischen Teilkompetenzen. Ein Bereich, bei dem dies insbesondere auffiel, ist die Teilkompetenz ‚Alternative Modelle'. Hier konnten Niveau I und II bei der Zuordnung nicht klar voneinander getrennt werden. Auf Niveau I werden alternative Modelle mit Unterschieden in den Modellobjekten begründet (Upmeier zu Belzen & Krüger, 2010); dies kann z. B. Größe, Material, Farbe oder die Dimensionalität (zwei- oder dreidimensionale Modelle) betreffen. Die Vorstellung, dass das Original die Herstellung verschiedener Modelle von etwas ermöglicht, das heißt die Modelle inhaltliche Foci auf verschiedene Aspekte des Originals abbilden, wird in Niveau II vertreten (Upmeier zu Belzen und Krüger, 2010) und ist in den entsprechenden Items so operationalisiert. Wenn die Dimensionalität als inhaltlicher Aspekt und nicht mehr als Eigenschaft der Modellobjekte verstanden wird, kommt es hier zu einer abweichenden Zuordnung. Die Abgrenzung zwischen den Niveaus wurde in der Skizzierung der relevanten Theorie auch in Bezug auf die Dimensionalität eines Modells deutlich gemacht und den Raterinnen und Ratern als Grundlage für die Zuordnung vorgelegt. Den Ausschlag für die Zuord-

nung als anderes Niveau gab demnach vermutlich ein individuell abweichendes Verständnis der theoretischen Grundlage.

Auch die Bereiche ‚Eigenschaften von Modellen', Niveau I und II, sowie ‚Testen von Modellen', Niveau II, waren wie beim Rating der Konstruktionsanleitung (Kapitel 3.1.4) für die Raterinnen und Rater schwer voneinander abzugrenzen und wurden teilweise als einer der beiden anderen Bereiche zugeordnet. Dies stützt die in Kapitel 3.1.4 formulierte Hypothese **H11**, dass diese Bereiche auch empirisch miteinander zusammenhängen könnten. Da sie unterschiedliche Facetten des Vergleichs von Modell und Original in den Blick nehmen und den Bezug (‚Eigenschaften von Modellen') bzw. Parallelen zwischen Modell und Original (‚Testen von Modellen') thematisieren, zeigte sich auch hier empirisch ein Zusammenhang, der bereits in der theoretischen Grundlage besteht. Entsprechend ist dieser Mangel an Übereinstimmung zwischen theoretischer und empirischer Zuordnung nicht als Hinweis auf eine mangelnde Repräsentation der theoretischen Fundierung in den einzelnen Items zu werten.

Insgesamt zeigte sich, dass ein tragfähiger Theoriebezug zwischen den einzelnen Items und dem Kompetenzmodell hergestellt werden kann. Daher kann angenommen werden, dass die einzelnen Items das Kompetenzmodell angemessen repräsentieren und für die empirische Strukturierung und Beschreibung von Modellkompetenz genutzt werden können.

3.3 Kognitive Prozesse bei der Itembearbeitung[8]

Die empirische Beschreibung von Modellkompetenz setzt außerdem voraus, dass die Items Schlüsse auf die Kompetenzausprägung von Schülerinnen und Schülern zulassen (Hartig et al., 2012). Aus diesem Grund stellt sich die Frage, inwiefern die kognitiven Prozesse, die zur Beantwortung eines Items führen, sich auf Modellkompetenz beziehen (vgl. Schmiemann, 2010). Zur Erhebung dieser Prozesse wurde die Methode des lauten Denkens eingesetzt (Kapitel 3.3.1). Die resultierenden Denkprotokolle von Schülerinnen und Schülern wurden nach der qualitativen Inhaltsanalyse (Mayring, 2010) von zwei Codiererinnen unabhängig ausgewertet und kriteriengeleitet auf die Validität der Items geprüft (Kapitel 3.3.2). Dadurch ergaben sich Befunde zur Validierung der Items, zu Bezügen innerhalb von Teilkompetenzen und zwischen ihnen sowie zum konkreten Umgang mit Modellen im Unterschied zu einem abstrakten Modellverständnis (Kapitel 3.3.3). Auf dieser Grundlage können sowohl Aussagen zu den Items als auch zum Konstrukt getroffen werden (Kapitel 3.3.4).

3.3.1 Lautes Denken

Borsboom et al. (2004, S. 1068) insistieren, dass die kognitiven Prozesse, die zur Beantwortung von Items führen, ausschließlich qualitativ erhoben werden können:

> "*No table of correlations, no matter how big, can be a substitute for knowledge of the processes that lead to item responses. The knowledge of such processes must be given by substantive psychological theory and cannot be based on methodological principles.*"

---

[8] Eine Zusammenfassung dieser Teilstudie findet sich in Terzer, Patzke und Upmeier zu Belzen (2012).

Zur Klärung der kognitiven Prozesse während der Bearbeitung der Items für die empirische Beschreibung von Modellkompetenz wurden entsprechend Denkprotokolle mit der Methode des lauten Denkens erhoben. Weidle und Wagner (1994) bewerten das laute Denken (Ericsson & Simon, 1993) als Methode, die „[…] am ehesten und vollständigsten die Möglichkeit (bietet), die im Individuum ablaufenden Kognitionen zu erfassen" (S. 83). Die befragte Person wird hierbei aufgefordert, alles laut auszusprechen, was sie denkt, während sie „mit einem Sachverhalt umgeht und eine Lösung für ein Problem sucht" (Städtler, 1998, S. 624; vgl. Weidle & Wagner, 1994). In Abgrenzung zu Interviews geht es bei dieser Methode darum, einen prozessorientierten Zugang zu gerade ablaufenden Kognitionen und mentalen Operationen zu finden, statt zu erfassen, was jemand über sich selbst denkt (Bilandzic, 2005; Weidler & Wagner, 1994). Demnach sollen Gedanken ausgesprochen statt reflektiert und theoretisiert werden (Bilandzic, 2005).

Kognitionspsychologische Grundlage für diese Methode ist das Prozessmodell nach Ericsson und Simon (1993), das zwischen zwei Gedächtnisformen, dem Langzeit- und dem Arbeitsgedächtnis, differenziert (Baddeley, 2002)[9]. Informationen im Arbeitsgedächtnis sind im Gegensatz zu Inhalten im Langzeitgedächtnis direkt für weitere Verarbeitungsschritte wie die Verbalisierung zugänglich. Beim lauten Denken sollen in Abgrenzung zur Introspektion, bei der es um die Beobachtung der eigenen Person und des eigenen Denkens geht, kognitive Prozesse unreflektiert und direkt während ihres Ablaufs wiedergegeben werden (Knoblich & Öllinger, 2006). Diese Form des lauten Denkens wird deshalb als *Online*-Verbalisierungsmethode bezeichnet, bei der die Inhalte des Arbeitsgedächtnisses unmittelbar zugänglich gemacht werden (Ericsson & Simon, 1993; Veenman, 2005). Dies ist ein entscheidender Vorteil im Vergleich

---

[9] Ericsson und Simon (1980, 1993) sprechen in diesem Zusammenhang noch vom Kurzzeit- statt Arbeitsgedächtnis. Dies hat jedoch keine Auswirkungen auf die hier beschriebene Methode.

zu retrospektiven Verbalisierungsmethoden wie einem Interview, in dem rekonstruiert wird, wie eine Aufgabe gelöst wurde. Beim lauten Denken entfallen Vergessens- und Inferenzprozesse, so dass es als weniger fehleranfällig gilt (Bannert, 2007).

Bei der hier vorgestellten Anwendung der Methode des lauten Denkens stand in Abgrenzung zur Selbsterklärung (Lind, Friege, Kleinschmidt & Sandmann, 2004) nicht im Vordergrund, wie Probandinnen und Probanden Kohärenzlücken in den Aufgaben füllen und über die Erweiterung ihres Wissens unter Verwendung des Vorwissens sowie der präsentierten Information reflektieren. Die Methode des lauten Denkens diente hier stattdessen dazu, nicht nur das Endergebnis einer Aufgabenbearbeitung, sondern vor allem die einzelnen Schritte der Informationsverarbeitung und die mit ihnen verbundenen Vorstellungen zu identifizieren (Bannert, 2007).

3.3.2 Datenerhebung und -auswertung

Da die Protokolle des lauten Denkens ausschließlich qualitativ und deskriptivstatistisch ausgewertet wurden, reichte für diese Testhefte ein einfaches Design aus (Anhang 6). Die Items wurden mit abwechselndem Schwierigkeitsgrad und zufälliger Zusammensetzung der Teilkompetenzen auf fünf Testhefte zu je neun Items verteilt. Diese Anzahl wurde nach Befunden der Pilotierungsstudie von Patzke (2010) ausgewählt. Von den befragten Schülerinnen und Schülern besuchten zwölf die Jahrgangsstufe 10 eines Gymnasiums und zehn die Jahrgangsstufe 7 einer Sekundarschule (hier fast ausschließlich Schülerinnen und Schüler mit Deutsch als Zweitsprache).

Für die Durchführung des lauten Denkens sind eine wertungsfreie, tolerante Atmosphäre und Anonymität wichtige Voraussetzungen (Bilandzic, 2005). Soziale Interaktion zwischen Versuchsleiterin und Proband wurde vermieden, um eine Beeinflussung der kognitiven Prozesse so weit wie möglich einzuschränken. Die Schülerin bzw. der Schüler führte deshalb eine Art Selbstgespräch, während die Versuchsleiterin im Hintergrund blieb. Die Instruktionen waren nondirektiv, z. B. „Denke bitte laut", „Rede weiter" (vgl. Ericsson & Simon, 1993). Fragen wie „Warum?" oder „Was meinst du damit?" wurden vermieden,

um keine Reflexionsprozesse auszulösen (vgl. Bilandzic, 2005). Zu den Aufgaben wurden keine Erklärungen gegeben, um auch diesen Einfluss zu vermeiden. Um das laute Denken zu routinisieren und die Hauptaufmerksamkeit auf die Primäraufgabe legen zu können, wurde ein Aufwärmtraining mit Übungsaufgaben durchgeführt (vgl. Bilandzic, 2005).

Bereits bei der Entwicklung der Items wurde z. B. dadurch, dass Schülerantworten auf eine offene Version der Items für die Formulierung der Antwortmöglichkeiten verwendet und sämtliche Schlüsselwörter darin im Itemstamm ergänzt wurden, eine möglichst hohe Validität angestrebt (Terzer et al., angen.). Darüber hinaus wurde durch eine Distraktorenanalyse geprüft, inwiefern die Distraktoren möglichst gleich plausibel waren, um einem Vorgehen nach dem Ausschlussprinzip vorzubeugen. Aus diesem Grund wurde das logische Schließen nicht explizit erhoben.

Die Durchführung der Methode des lauten Denkens wurde an Schülerinnen und Schülern der Jahrgangsstufe 10 (Gymnasium und Sekundarschule) pilotiert (Patzke, 2010), um ein Vorgehen zu ermitteln, das zu möglichst umfassenden Daten zu den Vorstellungen führt, die durch die Aufgabe bei den Schülerinnen und Schülern aktiviert werden. Es zeigte sich, dass die meisten auswertbaren Aussagen gewonnen werden konnten, wenn in einem ersten Schritt Itemstamm und Instruktion vorgelegt wurden und in einem zweiten Schritt die Antwortmöglichkeiten.

Die so erhaltenen Denkprotokolle wurden nach der qualitativen Inhaltsanalyse (Mayring, 2010) ausgewertet. Die verwendeten Kategorien basieren auf dem System, das Grünkorn, Upmeier zu Belzen und Krüger (in Vorb.) zur Auswertung von Aufgaben zur Modellkompetenz im offenen Antwortformat entwickelt haben. Dieses beinhaltet für die Teilkompetenzen ‚Alternative Modelle' sowie ‚Testen' und ‚Ändern von Modellen' zusätzlich zum Kompetenzmodell ein initiales Niveau (Grünkorn, Upmeier zu Belzen & Krüger, 2011; Grünkorn & Krüger, 2012). Schülerinnen und Schüler formulieren in diesen beiden Kategorien, dass Modelle nicht geändert werden müssen bzw. dass es keine alternativen Modelle gibt.

Der Codierleitfaden von Grünkorn et al. (in Vorb.) wurde zur Auswertung des hier vorliegenden Datentyps adaptiert und induktiv für den hier verwendeten Codierleitfaden (Anhang 7) ergänzt. Die neu definierten Kategorien beziehen sich auf die Beschreibung des Modellobjekts, die Bewertung des Modells, die Relevanz einer Antwortmöglichkeit für die Fragestellung, die Plausibilität einer Antwortmöglichkeit, die Aktivierung von Vorwissen sowie Ja/Nein-Antworten. Sämtliche Denkprotokolle wurden mit diesem Leitfaden von zwei wissenschaftlichen Mitarbeiterinnen aus dem Forschungsbereich Modellkompetenz unabhängig voneinander codiert. Die Übereinstimmung, in welche Kategorie die Aussagen eingeordnet wurden, und somit die Reliabilität der Codierung wurde über die Berechnung von Cohens κ geprüft (Wirtz & Caspar, 2002).

Ein Item wurde auf der Grundlage der doppelten Codierung dann als valide eingestuft, wenn folgende Kriterien erfüllt waren (Terzer et al., 2012):

- Die Schülerinnen und Schüler formulieren mit Bezug auf das Item Vorstellungen in der angesteuerten Teilkompetenz und Niveaustufe.
- Im Sinne einer Modellkompetenz fachlich angemessene Überlegungen sind mit einer fachlich richtigen Beantwortung verbunden sowie aus wissenschaftlicher Sicht nicht angemessene Überlegungen mit einer fachlich falschen Antwort. Mangelndes oder zusätzliches Fachwissen zu den biologischen Inhalten hat keine Auswirkungen auf die Beurteilung der Überlegungen. Stattdessen bezieht sich diese ausschließlich auf die Qualität der Vorstellungen zu Modellen und Modellbildung.

Wenn Schülerinnen und Schüler im Sinne einer Modellkompetenz fachlich richtig argumentierten und eine falsche Antwort ankreuzten oder umgekehrt bzw. ausschließlich andere Teilkompetenzen und/oder Niveaustufen ansprachen, wurde die Aufgabe als nicht valide eingestuft, da die Lösung nicht vor dem Hintergrund von Modellkompetenz interpretiert werden kann. Dies ist zu deren empirischer Beschreibung jedoch notwendig.

3.3.3 Ergebnisse

Validierung der Items

Die Codierung kann mit einem Cohens Kappa von κ = 0.97 ($N_{codierte\ Schüleraussagen}$ = 505) als sehr reliabel eingeschätzt werden (Wirtz & Caspar, 2002). Dabei wurde deutlich, dass die Denkprotokolle der Schülerinnen und Schüler der Jahrgangsstufe 7 der Sekundarschule mit *M* = 3.4 codierten Aussagen je Aufgabe (*SD* = 1.4) deutlich weniger Anhaltspunkte für die Beantwortung der Fragestellungen boten als die der Schülerinnen und Schüler der Jahrgangsstufe 10 des Gymnasiums mit *M*=9.1 codierten Aussagen je Aufgabe (*SD* = 3.9). Aufgrund der großen sprachlichen Verständnisschwierigkeiten und der damit einhergehenden geringen Aussagekraft der Sekundarschul-Denkprotokolle beziehen sich die weitere Darstellung der Ergebnisse und deren Interpretation ausschließlich auf die Denkprotokolle der sechs Gymnasiastinnen und sechs Gymnasiasten (*N* = 12).

Von den 45 untersuchten MC-Items wurden 40 Items nach den oben genannten Kriterien (Kapitel 3.3.2) als valide eingeordnet. Abb. 14 zeigt in der grün markierten Diagonale die Übereinstimmung von angesteuerter und von den Schülerinnen und Schülern formulierter Teilkompetenz und Niveaustufe (für die Zahlenwerte siehe Anhang 8). Ohne die Antwortmöglichkeiten zu kennen, formulierten die Schülerinnen und Schüler in 50 % ihrer Antworten selbst Inhalte, die in den Antwortmöglichkeiten vorkommen.

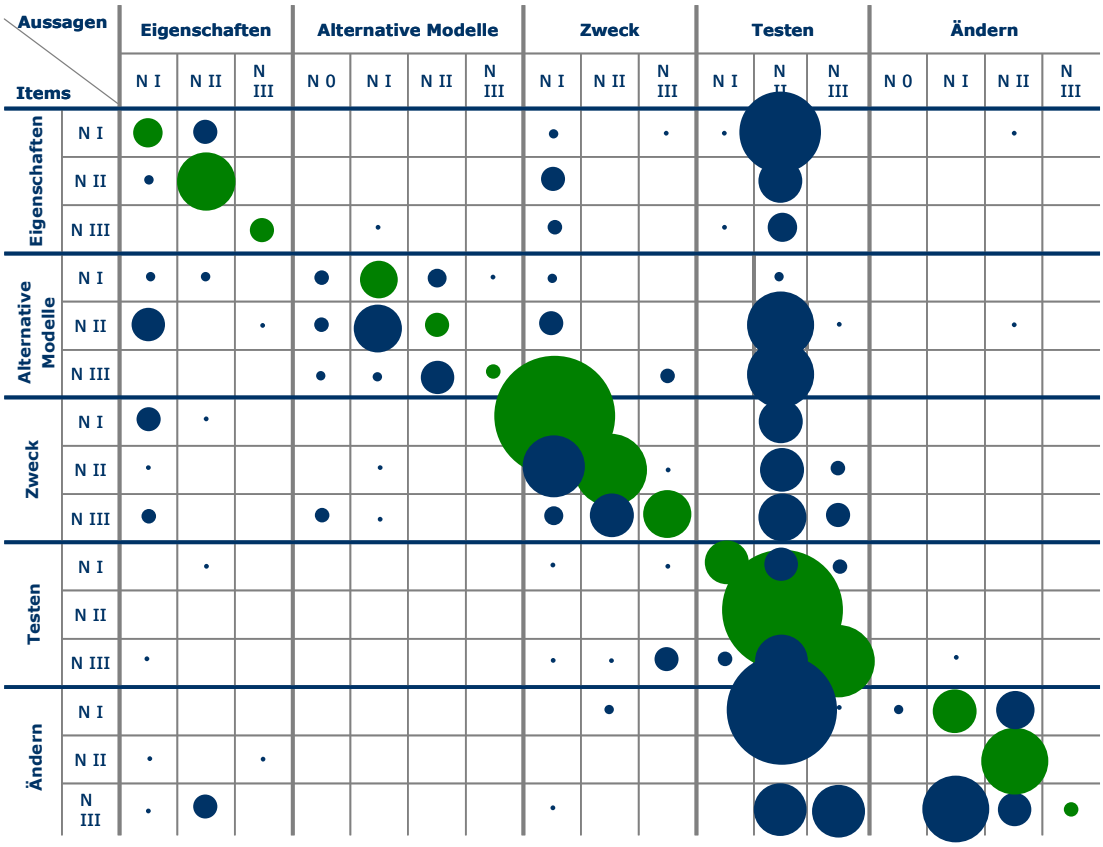


Abb. 14: Übersicht über die Verteilung der Schüleraussagen (*N* = 505). Die Zeilen entsprechen den angesteuerten Teilkompetenzen und Niveaustufen, die Spalten denen, in die die Schüleraussagen codiert wurden. Die Fläche der Kreise entspricht der Häufigkeit der Aussagen. Die grün markierten Kreise stellen die Aussagen dar, bei denen die mit dem Item angesteuerte Teilkompetenz und Niveaustufe mit den getroffenen Schüleraussagen übereinstimmten. Die blau markierten Kreise bilden Aussagen ab, die zu anderen Teilkompetenzen und/oder anderen Niveaustufen getroffen wurden.

Ein Beispiel für diese Übereinstimmung, die die Annahme der Validität eines Items legitimiert, sind die Vorstellungen, die Schüler G2 zu einem Item zum Erklären von Zusammenhängen mit einem Modell (‚Zweck von Modellen', Niveau II) formulierte:

> „Ich denke, damit [mit dem Speiseröhren-Modell] kann man den Weg der Nahrung im Zusammenhang mit den Muskeln und der Kontraktion,

> der Bewegung der Muskeln, erkennen. Wie sich die Muskeln bewegen, wann sie sich bewegen, wenn das Essen runter geht."

Er erklärte mit dem Modell demnach den Zusammenhang zwischen dem Transport der Nahrung und der Bewegung der Muskeln. Auch bei der Entscheidung für eine Antwortmöglichkeit dachte er in diesem Bereich:

> „Das mit dem Transport erscheint mir am logischsten, weil, wenn sie [Person aus dem Aufgabenstamm] die Speiseröhre zusammenzieht, dann steuert sie das selbst und dann kann sie sehen, wie gut sich die Nahrung bei bestimmten Bewegungen transportieren lässt."

Die Aufgabe ist demnach als valide einzustufen. Im Gegensatz zu dieser Aufgabenbearbeitung formulierte Schülerin G12 als eigene Lösung dazu, was mit einem Pflanzenzellenmodell veranschaulicht werden kann (‚Zweck von Modellen', Niveau I): „Ich würde (…) denken, dass er [Person aus dem Aufgabenstamm] die verschiedenen Teile der Pflanzenzelle zeigt und wie das Ganze aufgebaut ist, sozusagen 3D, dass man das noch einmal genauer sieht." Diese Lösung entspricht dem Attraktor. Bei der Auswahl einer Antwortmöglichkeit war sie jedoch irritiert:

> „Entweder würde ich sagen, würde er zeigen, wie eine Pflanzenzelle aufgebaut ist (Antwort 3), aber ich glaube, das kann man auch anhand des Bildes [Abbildung des Originals] sehen. Deswegen würde ich eher sagen, dass er die Größenverhältnisse zeigen möchte wie in einer Pflanzenzelle."

Der Grund, warum sie einen Distraktor ankreuzte, war demnach die mangelnde Plausibilität des Modells – es veranschaulicht ihrer Einschätzung nach nicht wesentlich mehr als das Original. Dass das Modell im Gegensatz zu einer mikroskopischen Abbildung der Pflanzenzelle den dreidimensionalen Aufbau einer Pflanzenzelle zeigt, nahm sie trotzdem wahr. Obwohl sie über Vorstellungen zum ‚Zweck von Modellen' im Bereich von Niveau I und somit die Kompetenz verfügt, die diese Aufgabe erfassen soll, wählte sie eine fachlich nicht ange-

messene Antwortmöglichkeit aus. Diese Aufgabe wurde deshalb als nicht valide aus dem Itempool ausgeschlossen.

Bezüge innerhalb von Teilkompetenzen und zwischen ihnen

Neben Aussagen zur Validität der Items erlaubt die Auswertung der Denkprotokolle außerdem Aussagen zu Bezügen innerhalb von Teilkompetenzen und zwischen ihnen. Zwei einzelne Teilkompetenzen wiesen Besonderheiten auf: Innerhalb der Teilkompetenz ‚Alternative Modelle' fiel auf, dass die Schülerinnen und Schüler häufig nicht nur Vorstellungen im Bereich eines Niveaus nannten, sondern verschiedene Niveaus verknüpften. Ein Beispiel hierfür ist die Antwort von Schülerin G1 auf eine Aufgabe zu ‚Alternativen Modellen', Niveau I:

> „Im Modell 1 sehe ich, wie die Knochen von der Haut überspannt sind. Im Modell 2 sehe ich, wie der Tyrannosaurus rex frontal zu mir gedreht ist und sogar Farbe bekommen hat, und in Modell 3 sehe ich den Tyrannosaurus rex mit Fell. In Modell 2 hat er kein Fell. Dann denke ich darüber nach, ob man irgendeinen Anhaltspunkt zur Farbe oder Fellform in den Knochen sieht, was nicht so ist. Deshalb glaube ich, dass Modell 2 und 3 sich deshalb voneinander unterscheiden, weil offensichtlich kein Stück Haut übrig geblieben ist und man nicht sehen kann, ob der Tyrannosaurus rex Fell hatte oder nicht."

Sie beschrieb hier verschiedene Eigenschaften, in denen die Modellobjekte sich unterscheiden (Niveau I), aber begründete alternative Modelle damit, dass man nicht wusste, wie das Original aussieht, und deshalb Annahmen treffen musste (Niveau III). Im gesamten Verlauf der Bearbeitung der Aufgabe sprach sie alle drei Niveaus an. Dies ist mit den Denkprotokollen anderer Schülerinnen und Schüler zu Aufgaben im Bereich ‚Alternative Modelle' vergleichbar.

Auch die Ergebnisse zur Teilkompetenz ‚Zweck von Modellen' wiesen eine Auffälligkeit auf. Hier trafen die Schülerinnen und Schüler vor allem Aussagen in Niveau I. Wenn sie Niveau II bzw. III ansprachen, nannten sie trotzdem häu-

fig auch Vorstellungen in Niveau I und/oder II (vgl. Abb. 14). So z. B. Schüler G4:

> „Vorhin hab ich gesagt, dass Flossenmodell 1 mir optisch mehr gefällt. Vom Prozess her ist aber doch Flossenmodell 2 maßstabsgetreuer, weil man hier den Vorgang gut sieht, dass die Flosse sich in die Richtung biegt, von wo der Druck kommt. Er kann wahrscheinlich ableiten, dass beim Fisch die Schwanzflosse nicht nur eine Hautfläche ist, sondern dass dort auch noch bewegliche Teile vorhanden sind und dass diese sich bei Druck eben wölbt, weil ein Fisch das braucht, um zu steuern."

Er beschrieb in diesem Zitat sowohl, dass das Modell etwas darstellt (Niveau I), als auch eine Voraussage über das Original, die man auf der Grundlage des Modells treffen kann (Niveau III).

Auch zwischen den Teilkompetenzen stellten die Schülerinnen und Schüler häufig Bezüge her (vgl. Abb. 14). Schülerin G10 antwortete auf den Aufgabenimpuls „Was kann man mit dem Roboter zeigen?" (‚Zweck von Modellen', Niveau I) z. B. zunächst mit einem Vergleich zwischen Modell und Original (‚Testen von Modellen', Niveau II). „Erst einmal fällt mir auf, dass der Roboter sehr menschenähnlich aussieht, auch die Bewegungen machen kann, wie es aussieht, und auch so weit eingeschränkt ist und so viel Freiheit hat wie der Mensch. " Daraus leitete sie ab, was man mit dem Roboter zeigen kann:

> „Man kann auf jeden Fall mit dem Roboter veranschaulichen, inwiefern sich der Fußballer oder der Mensch mit dem Ball und zu dem Ball bewegen kann, welche Freiheiten er hat, wie weit er z. B. seine Arme und seine Beine strecken kann."

Sie bezog sich außerdem auf die Teilkompetenz ‚Eigenschaften von Modellen': „Dadurch, dass die Proportionen auch stimmen, kommt das schon ziemlich nahe an den Fußballspieler heran." Dies bildete die Grundlage für ihre Bewertung des Modells: „Das ist zwar jetzt nicht die Aufgabe, aber ich würde sagen, das ist ein gutes Modell. Man kann es gut damit veranschaulichen." Auch in der Reflexion der Antwortmöglichkeiten stellte sie eine Verbindung zwischen

Parallelen zwischen Modell und Original und dem, was man mit dem Modell überhaupt veranschaulichen kann, her. Obwohl es sich um eine Aufgabe zum ‚Zweck von Modellen', Niveau I, handelte, in der es um die Veranschaulichung von Aspekten des Originals mit einem Modell geht, bezog sie die Teilkompetenzen ‚Testen von Modellen' sowie ‚Eigenschaften von Modellen' ein. Das Item wurde dennoch als valide eingestuft, da Schülerin G10 zwar auch Vorstellungen in anderen als der angesteuerten Teilkompetenz und Niveaustufe formulierte, sie aber bei der Lösung des Items vorwiegend dafür argumentierte, inwiefern das Modell etwas veranschaulicht.

Mit Blick auf Zusammenhänge zwischen den Teilkompetenzen fiel insgesamt auf, dass die Schülerinnen und Schüler das ‚Testen von Modellen', Niveau II, in allen Items zu anderen Teilkompetenzen und Niveaus ansprachen (Abb. 14). Diese Schüleraussagen betrafen vor allem die Kategorie ‚Vergleich von Modell und Original'. Bei der Bearbeitung der Items zum ‚Testen von Modellen', Niveau II, sprachen die Schülerinnen und Schüler dagegen nur über die angesteuerte Teilkompetenz und Niveaustufe, und zwar fast ausschließlich (20 von 25 Aussagen) über die Passung von Modell und Original. Auch bei Items zum ‚Ändern von Modellen', Niveau II, bezogen sich 12 von 16 Aussagen auf die Kategorie ‚mangelnde Passung', während nur fünf Aussagen ausschließlich die Kategorie ‚neue Erkenntnisse' betrafen.

Als weiterer Bereich wurde das experimentelle Testen von Hypothesen mit Modellen (‚Testen von Modellen', Niveau III) in einem anderen Bereich häufig zusätzlich zur Beantwortung der Items einer anderen Teilkompetenz herangezogen: Die Schüleraussagen zu Items im Bereich ‚Ändern von Modellen', Niveau III, beziehen sich häufig auch auf das ‚Testen von Modellen', Niveau III. Schülerin G9 bearbeitete z. B. eine Aufgabe zur Veränderung eines Regenwurm-Modells auf der Grundlage von Daten, die die zugrundeliegende Hypothese falsifizieren. Hierzu schlug sie eine veränderte Hypothese vor und implizierte damit eine Änderung des Modells: „Regenwürmer haben noch diese kleinen Borsten, die wahrscheinlich wie so eine Art Widerstand sind, womit sie sich nach vorne ziehen." Bei der Reflexion der Antwortmöglichkeiten schloss sie die Möglichkeit aus, dass „verschieden dicke Abschnitte" in das Modell auf-

genommen werden müssen, da sie es nicht für plausibel hielt, dass dies eine relevante Variable im Versuchsdesign ist: „Ich glaube, das spielt keine Rolle." Sie zog demnach in ihrer eigenen Lösung der Aufgabe die intendierte Fähigkeit heran, nutzte zur Reflexion der Antwortmöglichkeiten aber zusätzlich Fähigkeiten aus dem Bereich ‚Testen', Niveau III. Nur einer von sechs Schülerinnen und Schülern traf bei der Bearbeitung der Items zum ‚Ändern', Niveau III, keine Aussage über ‚Testen', Niveau III. Umgekehrt bezog sich keiner der Schülerinnen und Schüler, die Aufgaben zum ‚Testen', Niveau III, bearbeiteten, auf ‚Ändern', Niveau III.

Weitere Zusammenhänge zeigten sich bei den Items, die sich auf ‚Eigenschaften von Modellen', ‚Alternative Modelle' sowie ‚Zweck von Modellen', alle Niveaus, beziehen. In diesen Items trafen die Schülerinnen und Schüler häufig auch Aussagen darüber, inwiefern das Modell das Original veranschaulicht (Zweck von Modellen, Niveau I; vgl. Abb. 14). So beantwortete z. B. Schülerin G10 eine Aufgabe im Bereich ‚Alternative Modelle', Niveau III, folgendermaßen:

> „Da sich da [an einem Lungenfunktionsmodell] dann etwas bewegt, gibt es wahrscheinlich unterschiedliche Modelle zur Atmung. Einmal diesen Vorgang, um den darzustellen, wie es zum Beispiel vor der Atmung ist oder während man einatmet und der Brustkorb und die Lunge sind ausgedehnt."

In diesem Zitat kombinierte sie Vorstellungen zu ‚Alternativen Modellen', Niveau II, und Vorstellungen zum ‚Zweck von Modellen', Niveau I. Auch andere Schülerinnen und Schüler äußerten vergleichbare Kombinationen von Vorstellungen zum ‚Zweck von Modellen', Niveau I, mit anderen Niveaus zum ‚Zweck' oder Vorstellungen aus dem Bereich der ‚Kenntnisse über Modelle'.

Bei Aufgaben zum ‚Ändern von Modellen', Niveau III, war eine weitere Auffälligkeit festzustellen. Ein hier verwendetes Modell zum Insektenauge besteht aus Strohhalmen in einer Glasschale, durch das ein Bild an eine Wand projiziert wird. Schülerin G10 bewertete eine Antwortmöglichkeit zu diesem Item

z. B. so: „Das könnte ich mir vorstellen, wenn man noch mehr Strohhalme verwendet, dass dann die Strohhalme noch mehr Platz wegnehmen, weil das insgesamt von den Strohhalmwänden noch mehr Fläche ist.“ Statt die Hypothese zu verändern, die dem Modell zugrunde liegt und die durch die Daten aus dem entsprechenden Modellexperiment falsifiziert wurde, argumentierte sie dafür, das Modellobjekt so zu verändern, dass die Hypothese weiterhin angenommen werden kann, und bezog dies auf den Versuchsaufbau. Andere Schülerinnen und Schüler argumentierten über die mangelnde Passung zwischen Modell und Original, warum sie Hypothese nicht verwarfen. Diese Kombination aus Vorstellungen zum ‚Ändern von Modellen', Niveau I oder Niveau II, mit Vorstellungen zum ‚Testen von Modellen', Niveau III (Reflexion des Forschungsdesigns), trat bei drei von sechs Schülerinnen und Schülern auf.

Konkreter Umgang mit Modellen versus abstraktes Verständnis

Ein weiterer Befund betrifft Unterschiede darin, wie die Schülerinnen und Schüler eine eigene Lösung formulieren und welche Antwortmöglichkeiten sie dann auswählen. In den Denkprotokollen formulierten die Schülerinnen und Schüler siebenmal ohne Vorlage der Antwortmöglichkeiten eine Antwort, die dem Vorgehen in der angesteuerten Teilkompetenz und Niveaustufe entsprach, wählten aber dennoch einen Distraktor aus. Ein Beispiel dafür ist die Bearbeitung einer Aufgabe zum ‚Zweck von Modellen', Niveau III, von Schüler G8. In der Aufgabe wird ein Modellexperiment mit Fischmodellen gezeigt, die verschiedene Körperformen von Fischen darstellen. Sie werden von Gewichten an Schnüren durch einen mit Wasser gefüllten Blumenkasten gezogen. Schüler G8 nannte in seiner eigenen Lösung entsprechend dem Vorgehen in der angesteuerten Teilkompetenz und Niveaustufe eine Hypothese über Fische, die er aus dem gezeigten Modellexperiment ableitete: „Jana [Person in der Aufgabe] kann ableiten, dass die Jagdfische, die schnell sein müssen, oder schnelle Fische eher länglich und schmal geschnitten sind und langsame Fische dick sind.“ Die Formulierung dieser Hypothese ist im Sinne einer Modellkompetenz jedoch nicht angemessen, da sie nicht aus dem gezeigten Modellexperiment abgeleitet werden kann. Er entschied sich dann für den Distraktor „Die Geschwindigkeit von Fischen hängt vom Gewicht und von der Körperform

ab“. Es wird deutlich, dass er versuchte, anhand des Modells eine Hypothese abzuleiten und er diese Nutzung von Modellen somit abstrakt nachvollzieht. An diesem konkreten Beispiel gelang ihm dies aber nicht. Eine Antwortmöglichkeit lehnte er mit folgender Begründung ab: „Quatsch, eigentlich kann man das nicht so genau festlegen. Vielleicht gibt es irgendwelche Ausnahmen.“ Er unterschied demnach nicht, inwiefern es auf der Grundlage von Daten legitim ist, eine Hypothese abzuleiten, und inwiefern eine Hypothese gültig ist. Insgesamt waren die im Sinne einer Modellkompetenz fachlich nicht angemessenen Überlegungen hier mit einer fachlich falschen Beantwortung verbunden. Darüber hinaus zielen die Items nicht auf das abstrakte Verständnis von Modellen, sondern auf deren konkrete Anwendung ab. Somit ist die Aufgabe als valide einzustufen. Auch andere Bearbeitungen dieser Aufgabe weisen nicht auf eine mangelnde Validität hin.

3.3.4 Diskussion

Validierung der Items

Die Denkprotokolle der Schülerinnen und Schüler, die die Jahrgangsstufe 7 einer Sekundarschule besuchten und mit Deutsch als Zweitsprache Schwierigkeiten hatten, führten nicht zu aussagekräftigen Ergebnissen. Da sie vielfach zentrale Begriffe sprachlich nicht verstanden und somit die Grundlage für eine adäquate Bearbeitung der Aufgaben nicht gegeben war, sind diese Protokolle nicht sinnvoll auswertbar. Patzke (2010) konnte von Sekundarschülerinnen und -schülern der Jahrgangsstufe 10 auswertbare Denkprotokolle erheben. Das laute Denken eignet sich zur Itemvalidierung deshalb anscheinend erst in höheren Jahrgangsstufen der Sekundarschule. Da die Daten keine Hinweise auf die Validität der Items für diesen Teil der Zielgruppe liefern und große Verständnisschwierigkeiten beim lauten Denken deutlich wurden, wird von einem Einsatz der Items in der Sekundarschule abgesehen. Die Zielgruppe für die empirische Beschreibung von Modellkompetenz beschränkt sich deshalb auf Schülerinnen und Schüler des Gymnasiums, so dass Ergebnisse nicht über diese Gruppe hinaus generalisiert werden können (vgl. Rost, 2004). Hier erwies sich das laute Denken als geeignete Methode, um die Qualität der Operationalisierung zu prüfen, da Einblicke in die Vorstellungen, die Schülerinnen

und Schüler zur Beantwortung der Aufgaben heranziehen, gewonnen werden konnten. Außerdem lassen sich die Items auf die jeweilige theoretische Grundlage beziehen, weil die Schülerinnen und Schüler bei der Aufgabenbearbeitung nicht nur über Modelle reflektierten, sondern Vorstellungen in der jeweils angesteuerten Teilkompetenz und Niveaustufe formulierten.

Diese Vorstellungen lassen sich analog zu den Daten von Grosslight et al. (1991), Justi und Gilbert (2003), Crawford und Cullin (2005) sowie Grünkorn et al. (in Vorb.) kategorisieren. Somit scheint das Konstrukt, das mit diesen Items erfasst wird, an diese Untersuchungen zu Modellen anschlussfähig zu sein.

Bezüge innerhalb von Teilkompetenzen und zwischen ihnen

Innerhalb der Teilkompetenz ‚Alternative Modelle' streuten die Vorstellungen, die Schülerinnen und Schüler bei der Itembearbeitung formulierten, unabhängig vom Niveau des Items stark. Dies ist möglicherweise dadurch erklärbar, dass die Vorstellungen, auf die die Schülerinnen und Schüler zurückgriffen, unterschiedlich stark ad hoc entwickelt bzw. aus dem Langzeitgedächtnis heraus aktiviert wurden (Vosniadou, 2002). Die Schülerinnen und Schüler verfügten demnach möglicherweise in dieser Teilkompetenz über wenige Vorstellungen im Langzeitgedächtnis, so dass sie stattdessen spontan Vorstellungen entwickelten. Dies könnte eventuell darauf zurückgeführt werden, dass sie solche Vorstellungen selten brauchten, um Anforderungen einer Situation zu begegnen.

Dass in den höheren Niveaus der Teilkompetenz ‚Zweck von Modellen' auch Vorstellungen der darunter liegenden formuliert wurden, ist möglicherweise ein Indiz dafür, dass in dieser Teilkompetenz die Niveaus hierarchisch aufeinander aufbauen. Inwiefern dies der Fall ist, muss jedoch in einer längsschnittlichen Untersuchung geprüft werden (s. hierzu Patzke & Upmeier zu Belzen, 2011).

Zu den Items äußerten die Schülerinnen und Schüler nicht nur Vorstellungen aus den angesteuerten Bereichen. Diese Zusammenhänge wurden von Gross-

light et al. (1991), Justi und Gilbert (2003), Crawford und Cullin (2005) nicht beschrieben. Schwarz et al. (2009) weisen jedoch allgemein darauf hin, dass beim konkreten Umgang mit Modellen immer auf Vorstellungen aus verschiedenen Bereichen zurückgegriffen werden muss.

Zusammenhänge zwischen den Teilkompetenzen betreffen vor allem das ‚Testen von Modellen', Niveau II. Der Vergleich zwischen Modell und Original, der in dieses Niveau einzuordnen ist, wurde bei allen Items angesprochen. Dies lässt sich aus der Theorie heraus erklären: Der Vergleich von Modell und Original, den die hier befragten Schülerinnen und Schülern sehr häufig zogen, setzt beide in Beziehung. Dies entspricht der Sichtweise von Modellen als Modelle von bzw. für Originale (Mahr, 2008a). An dieser Stelle tritt dieses grundlegende Konzept bei der Bearbeitung der Items auf, das auch in den Ratings deutlich wurde (Kapitel 3.1.4, 3.2). Dies berührt nicht die Validität der Items, da eine Auseinandersetzung mit Modellen, wie sie in den Teilkompetenzen des Kompetenzmodells beschrieben wird, ohne diesen Bezug nicht sinnvoll möglich ist. Die Items, die ‚Testen', Niveau II, operationalisieren, bearbeiteten die Schülerinnen und Schüler dagegen vor allem unter Rückgriff auf Vorstellungen zur Passung von Modell und Original. Entsprechend ist kein besonderer Zusammenhang zwischen diesen Items und dem übrigen Itempool zu erwarten. Ein solcher liegt aber zu den Items zum ‚Ändern von Modellen', Niveau II, nahe, da sie unter Rückgriff auf die mangelnde Passung bearbeitet wurden. Dies spricht dafür, dass die Fähigkeit, die Passung zwischen Modell und Original zu beurteilen, für die Begründung einer Änderung des Modells aus der Herstellungsperspektive (Mahr, 2008a) benötigt wird. Bei der Strukturierung von Modellkompetenz ist deshalb zu vermuten, dass sich dieser Zusammenhang empirisch abbildet (**H13**; Abb. 15).

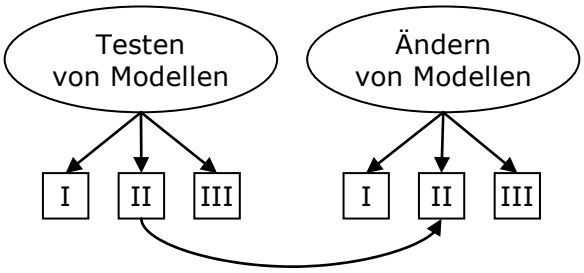


Abb. 15: Strukturmodell zu Hypothese **H13**. Die römischen Ziffern stehen für die Items der jeweiligen Niveaustufen.

Ein weiterer Zusammenhang zwischen Teilkompetenzen betrifft das ‚Testen' und ‚Ändern von Modellen', jeweils Niveau III. Da Fähigkeiten aus dem Bereich ‚Testen', Niveau III, für die Bearbeitung von Items aus dem Bereich ‚Ändern', Niveau III, genutzt wurden, aber nicht umgekehrt, ist ein Einfluss von ‚Testen', Niveau III, auf ‚Ändern', Niveau III, anzunehmen (**H14;** Abb. 16).

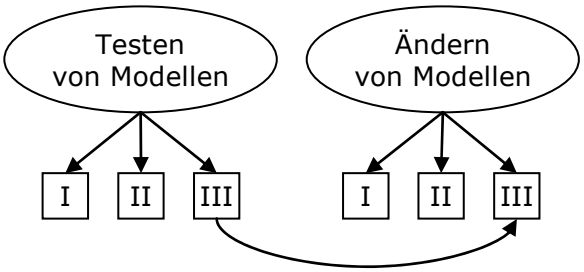


Abb. 16: Strukturmodell zu Hypothese **H14**. Die römischen Ziffern stehen für die Items der jeweiligen Niveaustufen.

Auch die Fähigkeiten, die den ‚Zweck von Modellen', Niveau I, betreffen, wurden zur Bearbeitung von Items zu anderen Bereichen des Kompetenzmodells herangezogen. Dies betrifft sowohl die ‚Kenntnisse über Modelle' als auch Niveau II und III beim ‚Zweck von Modellen', was einen Zusammenhang zwischen diesen Bereichen vermuten lässt (**H15a;** Abb. 17). Möglicherweise spricht dieser Befund sogar dafür, den ‚Zweck von Modellen' der Dimension ‚Kenntnisse über Modelle' statt der Modellbildung zuzuordnen (**H15b;** Abb. 18). Dies würde der Einordnung dieses Aspekts nach Leisner-Bodenthin (2006) entsprechen, die Vorstellungen zum ‚Zweck von Modellen' im deklara-

tiven Wissen und somit im Gegensatz zu Justi und Gilbert (2002) nicht im Modellbildungsprozess verortet.

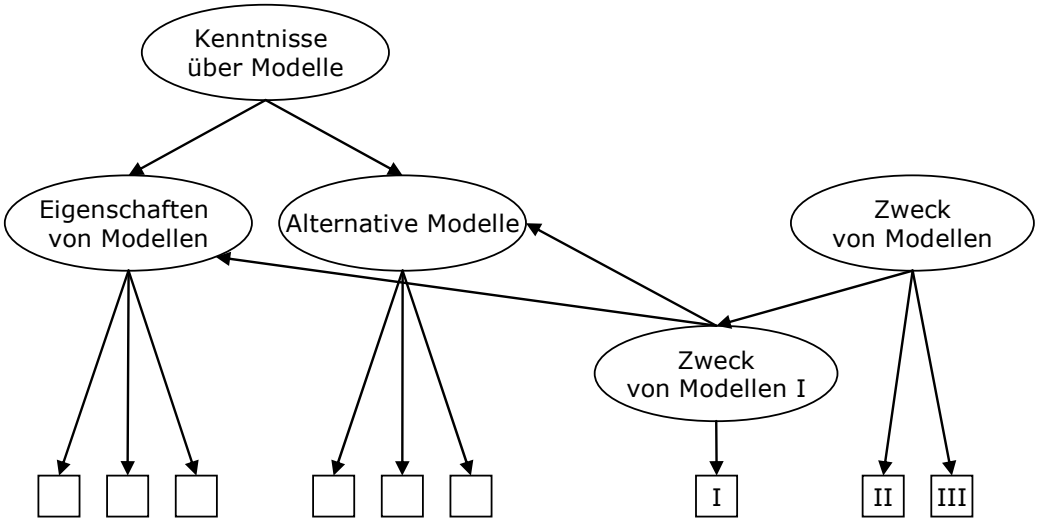


Abb. 17: Strukturmodell zu Hypothese **15a**. Die römischen Ziffern stehen für die Items der jeweiligen Niveaustufen.

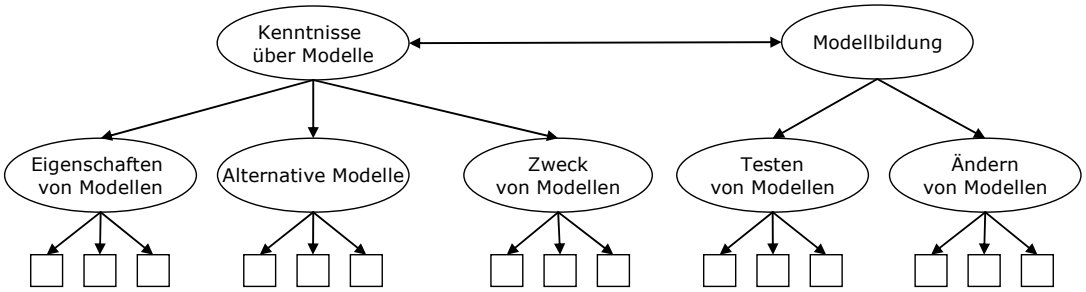


Abb. 18: Strukturmodell zu Hypothese **15b**.

Die Befunde, dass Schülerinnen und Schüler häufig nicht über methodisches Wissen über Ziele und Vorgehensweisen beim Experimentieren verfügen (Hammann et al., 2006), spiegelten sich in der Bearbeitung der Items zum ‚Ändern von Modellen', Niveau III, wider. Anstatt die Hypothese, die einem Modell zugrunde liegt, zu verändern, argumentierten die Schülerinnen und Schüler häufig, dass das Modellobjekt bzw. die Passung zwischen Modell und

Original mit Blick auf das Forschungsdesign verändert werden sollte. Hier spielten möglicherweise inhaltliche Eingangsüberzeugungen oder die Strategie des positiven Testens eine Rolle (Chinn & Brewer, 1993; Hammann et al., 2006). Da dies bei den befragten Schülerinnen und Schülern sehr häufig auftrat, sollte man dies bei der Förderung von Modellkompetenz in den Blick nehmen und grundsätzliche Strategien und Vorgehensweisen beim Experimentieren thematisieren (s. hierzu z. B. Chinn & Brewer, 1993; Hammann et al., 2006).

Konkreter Umgang mit Modellen versus abstraktes Verständnis

Der Befund, dass die Schülerinnen und Schüler in einigen Fällen über ein abstraktes Verständnis von Modellen verfügten, es aber in der konkreten Anwendung nicht umsetzen konnten, korrespondiert mit Ergebnissen einer Interventionsstudie von Leisner und Mikelskis (2004). Diese stellten fest, dass sie einen signifikanten Zuwachs im domänenübergreifenden Wissen über Modelle erzielen konnten, während sich ein weniger klares Bild für domänenspezifisches Wissen zu konkreten Modellen ergab. Urhahne et al. (2011) berichten, dass eine Vertrautheit mit verschiedenen Kontexten mit einem elaborierten Wissenschaftsverständnis einhergeht und ein Zusammenhang zwischen kontextunabhängigen und kontextspezifischen Vorstellungen besteht. Diese Befunde unterstützen die Bedenken von Sins et al. (2009), dass die kontextspezifische und die kontextunabhängige Erfassung von Vorstellungen zu Modellen zu unterschiedlichen Ergebnissen kommen kann. Ob damit die Validität kontextunabhängiger Instrumente grundsätzlich in Frage zu stellen ist, wie Sins et al. (2009) argumentieren, ist damit nicht gesagt. Es erscheint stattdessen plausibel, diese Facetten getrennt voneinander zu betrachten und explizit aufeinander zu beziehen. Schwarz et al. (2009) werfen in diesem Zusammenhang die Frage auf, inwieweit Schülerinnen und Schüler von bekannten auf unbekannte Kontexte transferieren, und damit, wie diese Facetten miteinander zusammenhängen.

3.4 Fazit zur Operationalisierung des Kompetenzmodells

Bereits im Verlauf der Instrumententwicklung wurden Verbindungen zur theoretischen Grundlage hergestellt und somit Aspekte der Validierung einbezogen. Die Konstruktionsanleitung, die als Grundlage für die Itementwicklung formuliert wurde, ließ sich dem Kompetenzmodell zuverlässig zuordnen und operationalisiert es demnach angemessen. Die hier gewählte Vorgehensweise (vgl. Terzer et al., angen.) erscheint deshalb als geeignet für die Test- und Itemkonstruktion auf der Grundlage von Kompetenzmodellen.

Auch wenn damit die entwickelten Items für die anvisierte Zielgruppe insgesamt eher leicht sind und mehr schwere Items sowie eine höhere Varianz und eine höhere Reliabilität wünschenswert wären (vgl. Kapitel 3.1.6), sind die entwickelten Items aus psychometrischer Sicht für die empirische Beschreibung von Modellkompetenz geeignet. Da auch die einzelnen Items dem Kompetenzmodell mit einer guten Übereinstimmung zugeordnet wurden, kann angenommen werden, dass sie es angemessen repräsentieren. Vorstellungen, die Schülerinnen und Schüler bei der Bearbeitung der Items formulierten, korrespondieren mit anderen Untersuchungen zu Modellen und lassen sich auf Modellkompetenz beziehen. Deshalb ist es legitim, die Beantwortung der Items als Ausprägung von Modellkompetenz zu interpretieren. Demnach ermöglicht die Operationalisierung von Modellkompetenz in MC-Items weitgehend die Erfassung von Modellkompetenz und kann für deren empirische Beschreibung genutzt werden.

4 Empirische Beschreibung von Modellkompetenz

Nationale und internationale Studien (z. B. Grosslight et al., 1991; Treagust et al., 2002; Trier & Upmeier zu Belzen, 2009) zeigen auf, dass Schülerinnen und Schülern nicht über eine umfassend ausgeprägte Modellkompetenz verfügen – die Frage ist, wie Schülerinnen und Schüler diese erwerben können. Erste Antworten auf diese Frage geben empirische Befunde zur Strukturierung und Graduierung von Modellkompetenz: Sie prüfen das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) als Referenzsystem für das professionelle Handeln von Lehrkräften (Kapitel 1). Als Anknüpfungspunkt für die Förderung von Modellkompetenz ist für Lehrerinnen und Lehrer außerdem Wissen über Lernvoraussetzungen, und zwar verfügbare Kompetenzen von Schülerinnen und Schülern, relevant. Darüber hinaus informieren Befunde zu Beziehungen von Modellkompetenz zu anderen Konstrukten (wie z. B. allgemeinen kognitiven Fähigkeiten) über die Erlernbarkeit und Domänenspezifität von Modellkompetenz. Sie geben Hinweise, ob etwa in Hinblick auf Geschlecht oder Wissenschaftsverständnis eine differenzierte Förderung erfolgen sollte. Deshalb ist die empirische Strukturierung, Graduierung und Erfassung von Modellkompetenz mithilfe des evaluierten Tests Gegenstand dieses Kapitels (Abb. 19).

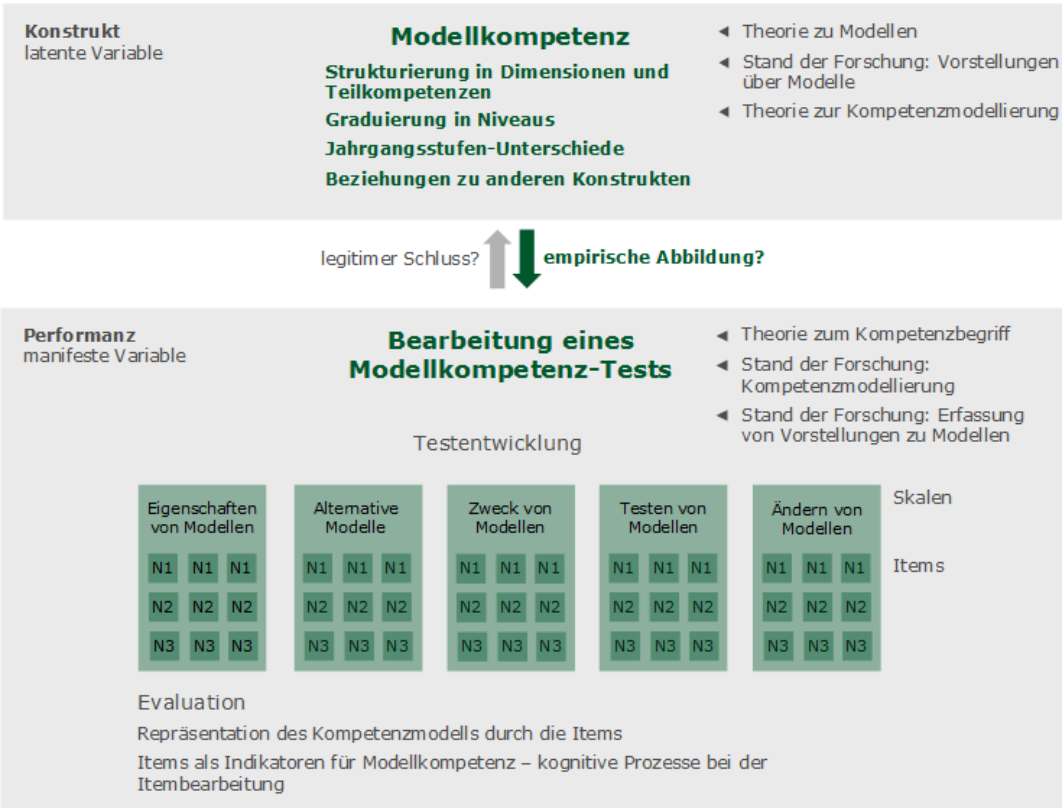


Abb. 19: Konzeption des Projekts – empirische Beschreibung von Modellkompetenz mithilfe des entwickelten, evaluierten Tests.

Für die empirische Beschreibung von Modellkompetenz wurde ein entsprechendes Untersuchungsdesign entwickelt (Kapitel 4.1): Die Stichprobe wurde aus der Population der Jahrgangsstufen 7 bis 10 des Gymnasiums gezogen, da die entsprechenden Bildungsstandards Kompetenzen zum MSA beschreiben (KMK, 2005). Schülerinnen und Schüler der Sekundarschule hatten große sprachliche Verständnisschwierigkeiten, so dass die Validität des in diesem Projekt entwickelten Tests für diese Gruppe fraglich ist (Kapitel 3.3). Neben den Modellkompetenzitems bearbeiteten die Schülerinnen und Schüler ergänzende Tests zu anderen Konstrukten wie z. B. allgemeinen kognitiven Fähigkeiten. Da eine Bearbeitung aller Items zeitlich nicht möglich wäre, wurde ein Multi-Matrix-Design gewählt, in dem jede Schülerin und jeder Schüler nur einen Teil der Items bearbeitet. Die zentrale methodische Grundlage bilden Analysen, die auf IRT-Modellen basieren (Kapitel 4.2). Auf dieser Basis wurden Ergebnisse zur empirischen Beschreibung von Modellkompetenz gewonnen

(Kapitel 4.3). Diese werden bezogen auf die jeweilige theoretische Fundierung diskutiert, um Modellkompetenz für eine effektive Förderung im Biologieunterricht zu beschreiben (Kapitel 4.4).

4.1 Untersuchungsdesign

Stichprobe

In Passung zu den Bildungsstandards für den Mittleren Schulabschluss im Unterrichtsfach Biologie (KMK, 2005) und in Anlehnung an das Vorgehen in bik-Teilprojekten (Kapitel 2.1.2) wurden Schülerinnen und Schüler der Jahrgangsstufen 7 bis 10 in Berlin befragt ($n_{Jgst.\ 7}$ = 313; $n_{Jgst.\ 8}$ = 293; $n_{Jgst.\ 9}$ = 284; $n_{Jgst.\ 10}$ = 288; N = 1178), um in einer Querschnittstudie die Variation der Kompetenz über mehrere Jahrgangsstufen hinweg erheben zu können. Da die Validität der Items in der Sekundarschule infrage steht (Kapitel 3.3), wurde ausschließlich das Gymnasium einbezogen. Da die Itemerprobung sowohl in der Realschule als auch im Gymnasium durchgeführt wurde, werden in Anhang 9 die Itemkennwerte für die hier betrachtete Stichprobe berichtet. Die befragten 626 Schülerinnen (53.14 %) und 543 Schüler (46.10 %; neun fehlende Angaben zum Geschlecht) waren 9 bis 19 Jahre alt (M = 14.3; SD = 1.4).

Ein Querschnitt erlaubt Aussagen auf Populationsebene über die interindividuelle Variation. Die Frage, ob das Kompetenzmodell auch ein Kompetenzentwicklungsmodell darstellt und intraindividuelle Variation, d. h. die Entwicklung der Kompetenz, abbildet, wird in weiteren Untersuchungen mit längsschnittlichen Datenerhebungen (Patzke & Upmeier zu Belzen, 2011) geklärt (vgl. Borsboom et al., 2004).

Instrumente

Eine Bearbeitung aller 40 Items würde etwa 100 Minuten in Anspruch nehmen und darüber hinaus nach Beobachtungen während der Itemerprobung (Kapitel 3.1.6) die Konzentration und Motivation der Schülerinnen und Schüler überanstrengen. Indem die Testhefte nach einem *balanced incomplete block*

design (BIDB) zusammengesetzt wurden, konnte eine vollständige Kovarianzmatrix erzeugt werden, obwohl die Schülerinnen und Schüler nicht jedes Item bearbeiten mussten (Giesbrecht & Gumpertz, 2004). Dazu wurden alle Items je Teilkompetenz und Niveau als Block gebündelt und über Testhefte rotiert (Tab. 12). Dieser Typ von Multi-Matrix-Design ist dadurch gekennzeichnet, dass

- jedes Testheft nur eine Teilmenge der Itemblöcke enthält und nicht alle möglichen Blockkombinationen auftreten (*incomplete*),
- kein Itemblock öfter als einmal je Testheft enthalten ist (*binary*) sowie
- alle Paare von Itemblöcken gleich oft vorkommen, so dass alle geschätzten Itemunterschiede eine gemeinsame Varianz haben und eine vollständige Kovarianzmatrix entsteht (*balanced*) (Giesbrecht & Gumpertz, 2004).

Tab. 12: Multi-Matrix-Testheft Design – (15, 3, 1)-BIBD (Colbourn & Dinitz, 1996).

	Testheft																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Position 1	Z1	T1	E2	E1	Ä2	E1	E1	A1	Ä1	Z2	A1	A3	Ä3	Z1	Ä1	Z2	Z1	Ä3	Z3
Position 2	A1	E1	A2	Z2	E1	Z3	T3	T1	A1	Ä2	T2	A1	Z3	A2	Z1	E3	T2	Z1	T3
Position 3	E1	Ä1	E1	T2	E3	A3	Ä3	E2	A2	A1	E3	T3	A1	T1	E2	Z1	Ä2	A3	Z1

Fortsetzung Tab. 12.

	Testheft															
	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Position 1	T1	T2	T3	T1	Z3	A3	Ä2	E3	E2	T2	Ä2	E3	A2	T3	A2	A2
Position 2	A3	T1	Ä2	E3	Ä1	Ä1	Ä3	Ä1	Z2	E2	A3	E2	Ä3	T2	Z3	E3
Position 3	Z2	Z3	T1	Ä3	Z2	T2	Ä1	T3	T3	Ä3	E2	Z3	Z2	A2	Ä2	A3

15 Blöcke werden in diesem Design so auf drei Testheftpositionen verteilt, dass jedes Paar von Blöcken einmal auftritt. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen.

Da jedes Paar von Blöcken in den 35 Testheften einmal auftritt, beträgt die Abdeckung der Kovarianzmatrix 0.029 (1/35). Um den Effekt der Blockposition zu minimieren, kommen jeder Block, jede Teilkompetenz und jede Niveaustufe an jeder Position gleich oft vor[10]. Indem die Testhefte im *spiraling*-Verfahren (Frey, Hartig & Rupp, 2009) auf die Schulklassen verteilt wurden, wurde die proportional geschichtete Klumpenstichprobe (Bortz, 2005) so weit wie möglich gemischt.

Kompetenzen werden kontextspezifisch in domänenspezifischen Situationen erworben und können somit durch Interventionen beeinflusst werden (Kapitel 2.1.1). Da allgemeine kognitive Fähigkeiten im Vergleich dazu viel stabiler sind, argumentieren Koeppen et al. (2008), dass bei der Konstruktion von Kompetenzmodellen die Verbindung der spezifischen Kompetenz und allgemeinen kognitiven Fähigkeiten empirisch betrachtet werden sollte (vgl. Kapitel 2.4.2). Zur Erhebung dieser Kontrollvariable wurde der KFT 4-12R (Skala N2 non-verbal; Heller & Perleth, 2000) eingesetzt.

Als weitere Kontrollvariable wurde das Wissenschaftsverständnis erhoben, da es über die theoretische Fundierung der Niveaus (Kapitel 2.3.3) theoriegeleitet besonders eng mit Modellkompetenz zusammenhängt (vgl. Kapitel 2.4.2). Hierfür wurde ein *Nature of Science*-Fragebogen von Urhahne et al. (2008) eingesetzt. Die Skalen dieses Fragebogens, die in Bezug auf Modellkompetenz

---

[10] Die Datenerhebung fand kombiniert mit zwei weiteren Aufgabenformaten statt (FC-Items, Krell & Krüger, 2011; Aufgaben in offenem Antwortformat, Grünkorn & Krüger, 2012), die das gleiche Testheftdesign nutzten. Die Testhefttteile wurden getrennt geheftet und ausgeteilt. Sie können einander über Schülernummern zugeordnet werden, so dass jeder Schüler und jede Schülerin Aufgaben zu den gleichen Teilkompetenzen in verschiedenen Antwortformaten bearbeitete. Dies ermöglicht den Vergleich der Formate. Um eine gegenseitige Beeinflussung zu vermeiden, wurden den Schülerinnen und Schülern zuerst die Aufgaben in offenem Antwortformat, als zweites die MC-Items und zuletzt die FC-Items vorgelegt.

relevant sind, sind ‚Herkunft' (5 Items), ‚Sicherheit' (7 Items), ‚Entwicklung' (8 Items) sowie ‚Rechtfertigung' (9 Items). In der hier vorgestellten Untersuchung wurden die Items nicht sortiert nach diesen Skalen, sondern zufällig in einem Fragebogen angeordnet. Die interne Konsistenz dieser Skalen liegt nach Urhahne et al. (2008) bei $.66 < \alpha < .71$.

Die MC-Items beinhalten im Vergleich zu z. B. Aufgaben in offenem Antwortformat viel Text, der für eine adäquate Itembearbeitung gelesen und verstanden werden muss. Deshalb ist ein Zusammenhang zu Lesefähigkeiten zu erwarten, die mit dem LGVT 6-12 (Schneider, Schlagmüller & Ennemoser, 2007) erhoben wurden, um einen Einfluss dieser Variable kontrollieren zu können. Weitere Kontrollvariablen waren Alter und Geschlecht der Schülerinnen und Schüler.

Den Vorgaben in den jeweiligen Manualen folgend (Heller & Perleth, 2000; Schneider et al., 2007) gingen jeweils die T-Werte von LGVT und KFT in die weiteren Berechnungen ein. Für die Items zum Wissenschaftsverständnis wurden in Anlehnung an die gängige Verwendung dieses Fragebogens (Urhahne et al., 2008) Mittelwerte gebildet.

Datenerhebung

Die Durchführung wurde für die Testleiterin bzw. den Testleiter in einer Anleitung standardisiert, die Informationen zur Vorstellung des Projekts, dem Testmaterial, der gewünschten Bearbeitung in Einzelarbeit sowie zum zeitlichen und organisatorischen Ablauf sowie standardisierte Antworten auf häufige Rückfragen von Schülerinnen und Schülern umfasste (vgl. Kapitel 3.1.6). Darüber hinaus protokollierte die Testleiterin bzw. der Testleiter Störungen während der Testphase, die sich auf die Bearbeitung des Tests auswirken könnten. Schülerinnen und Schülern, die früh mit der Bearbeitung der Testhefte fertig waren, erhielten zur Beschäftigung ein Rätsel, um Unruhe in der Klasse zu vermeiden. Die Testhefte waren nummeriert, so dass die Befragung vollständig anonym durchgeführt werden konnte.

4.2 Analysen

Einen Forschungsbereich der Kompetenzmodellierung bildet die Entwicklung entsprechender psychometrischer Modelle, die beschreiben, wie ein Ergebnis im Sinne des jeweiligen Kompetenzmodells zu interpretieren ist (Klieme & Leutner, 2006; Koeppen et al., 2008; Kapitel 2.1.1). Für die Auswahl eines psychometrischen Modells nennt Hartig (2008) Dimensionalitätsannahmen, das Skalenniveau der latenten Variable sowie Annahmen darüber, inwiefern z. B. ausgeprägte Fähigkeiten in einer Teilkompetenz mangelnde Fähigkeiten in einer anderen kompensieren können.

Insbesondere wenn Multi-Matrix-Designs eingesetzt werden und folglich keine vollständigen Datensätze vorliegen (Kapitel 4.1), werden häufig IRT-Modelle eingesetzt (z. B. in bik-Projekten, vgl. Kapitel 2.1.2). Hier existieren sowohl ein- als auch mehrdimensionale psychometrische Modelle, die einen Vergleich miteinander erlauben. Somit sind Modelle der IRT dafür geeignet, ein Kompetenzmodell auf seine empirische Dimensionalität zu prüfen. Auf Ebene der latenten Variable werden *kontinuierliche* Variablen wie z. B. allgemeine kognitive Fähigkeiten mit *quantitativ* unterschiedlichen Ausprägungen von *kategorialen* Variablen wie z. B. Einstellungstypen mit *qualitativ* unterschiedlichen Ausprägungen unterschieden. Für kontinuierliche Variablen wie Modellkompetenz werden sog. *Latent-Trait*-Modelle herangezogen (Moosbrugger, 2012). Diese beschreiben die Itembearbeitung als Interaktion von Personen- und Itemmerkmalen (Embretson & Reise, 2000; Reckase, 1997). Da Modellkompetenz theoriegemäß keine aufeinander aufbauenden Schritte beinhaltet, kann angenommen werden, dass unterschiedlich ausgeprägte Teilkompetenzen einander ausgleichen können. Somit ist die Verwendung kompensatorischer Modelle wie multidimensionaler IRT-Modelle adäquat. IRT-Modelle sind damit für die angestrebten Analysen angemessen und somit wesentliche methodische Grundlage des hier vorgestellten Projekts (Kapitel 4.2.1).

Auch wenn IRT-Modelle mit fehlenden Daten umgehen können, erfordern Analysen mit Hintergrundmodellen vollständige Datensätze für die Variablen, die im Hintergrund zur genaueren Schätzung der Personenfähigkeit einbezogen werden sollen. Dies ist z. B. dann notwendig, wenn Variablen zu Personenfä-

higkeiten in Bezug gesetzt werden sollen, da über das Hintergrundmodell der Einfluss dieser Variablen aus den Personenfähigkeiten auspartialisiert wird. Hierfür werden vollständige Daten für diese Variablen benötigt (Wu et al., 2007). In diesem Projekt betrifft dies die Variablen Wissenschaftsverständnis, allgemeine kognitive Fähigkeiten, Lesefähigkeiten, Geschlecht sowie Schulnoten. Aus diesem Grund wird zunächst der Umgang mit fehlenden Werten berichtet (Kapitel 4.2.2).

Da Modellkompetenz theoriegeleitet in fünf Teilkompetenzen und zwei Dimensionen strukturiert ist (vgl. Kapitel 2.4.1, 2.5), ist es notwendig, dass diese Annahmen zur Struktur in entsprechenden psychometrischen Modellen abgebildet werden. Dies ist mit IRT-Modellen möglich. Entsprechend wurde ein eindimensionales Modell (**H0**) mit einem zwei- (**H1**) und einem fünfdimensionalen (**H2**) verglichen (Kapitel 4.2.3). Das Strukturmodell, das in diesen Analysen die beste Passung mit den Daten aufwies, wurde für die weiteren Berechnungen herangezogen: Die darüber berechneten Itemschwierigkeiten wurden zur Prüfung der Graduierung von Modellkompetenz genutzt (Kapitel 4.2.4). Die in diesem Modell geschätzten Personenfähigkeiten bildeten die Grundlage für den Vergleich der Jahrgangsstufen 7 bis 10 (Kapitel 4.2.5). Auch die Zusammenhänge von Modellkompetenz mit anderen Konstrukten wie allgemeinen kognitiven Fähigkeiten oder Wissenschaftsverständnis wurden mithilfe der Personenfähigkeiten berechnet (Kapitel 4.2.6).

4.2.1 IRT-Modellierung

Die Interaktion von Personen- und Itemmerkmalen wird für eindimensionale Konstrukte häufig durch das Rasch-Modell als Lösungswahrscheinlichkeit für ein Item beschrieben. Es hat gegenüber anderen *Latent-Trait*-Modellen wie

dem 2PL- oder 3PL-Modell[11] den Vorteil, dass individuelle Personenfähigkeiten kriteriumsorientiert durch ihre Abstände zu Itemschwierigkeiten interpretiert werden können. Dies ist darin begründet, dass sowohl Itemschwierigkeiten als auch Personenfähigkeiten als Lösungswahrscheinlichkeiten berechnet und auf einer gemeinsamen Skala angegeben werden (*joint scale*; Moosbrugger, 2012). Hierfür wird die Itemschwierigkeit auf Grundlage der Anzahl der Personen geschätzt, die das jeweilige Item gelöst haben. Zur Schätzung der Personenfähigkeiten werden die Anzahl und die Schwierigkeiten der gelösten Items herangezogen. Auf diese Weise wird die verfügbare Kompetenz der Personen (latente Variable, hier Modellkompetenz) aus der beobachteten Performanz im Test (manifeste Variable, hier Modellkompetenz-Items bzw. deren Bearbeitung) rekonstruiert (Rauch & Hartig, 2012). Außerdem können auf der Grundlage dieser psychometrischen Modelle Kompetenzausprägungen auf spezifische situative Anforderungen bezogen werden, die in Itemmerkmalen abgebildet sind (Hartig, 2008; Rauch & Hartig, 2012). Da in die Schätzung der Itemschwierigkeiten nur die Lösungshäufigkeiten und in die Schätzung der Personenfähigkeiten die Anzahlen der gelösten Items je Person sowie die zuvor geschätzten Itemschwierigkeiten eingehen, können auch unvollständige Datensätze mit IRT-Modellen analysiert werden (Rost, 2004).

Die Abhängigkeit der geschätzten Personenfähigkeit von der Lösungswahrscheinlichkeit der Person für ein Item wird in einer ICC abgebildet (Embretson & Reise, 2000; Moosbrugger, 2012; Abb. 20). Die Einheit der hierzu verwendeten Skala bilden sog. *logits.* Diese sind der natürliche Logarithmus des Wettquotienten (*odd*) aus der Lösungswahrscheinlichkeit und der Gegenwahrscheinlichkeit, d. h. der Wahrscheinlichkeit, das Item nicht zu lösen (Embretson & Reise, 2000; Rauch & Hartig, 2012). Die Metrik dieser Skala wird durch

---

[11] Das 2PL-Modell (Zweiparamter-Logistisches-Modell) erweitert das Rasch-Modell um die Itemtrennschärfe als zweiten Itemparameter neben der Itemschwierigkeit. Im 3PL-Modell (Dreiparameter-Logistisches-Modell) wird darüber hinaus als dritter Itemparameter ein Rateparameter eingeführt (Moosbrugger, 2012).

die Restringierung der durchschnittlichen Itemschwierigkeit oder der durchschnittlichen Personenfähigkeit auf Null festgelegt (Rauch & Hartig, 2012). Wenn Informationen über die Items gewonnen oder Strukturmodelle verglichen werden sollen, wird der Mittelwert der Personenfähigkeiten auf Null festgesetzt (fallzentrierte Analysen). Für Analysen der Personenfähigkeiten wird der Mittelwert der Itemschwierigkeiten auf Null fixiert (itemzentrierte Analysen). Diese sind im Rasch-Modell als Wendepunkt der jeweiligen ICC definiert, an dem die Lösungswahrscheinlichkeit für das Item bei 50 % liegt (Moosbrugger, 2012; Abb. 20).

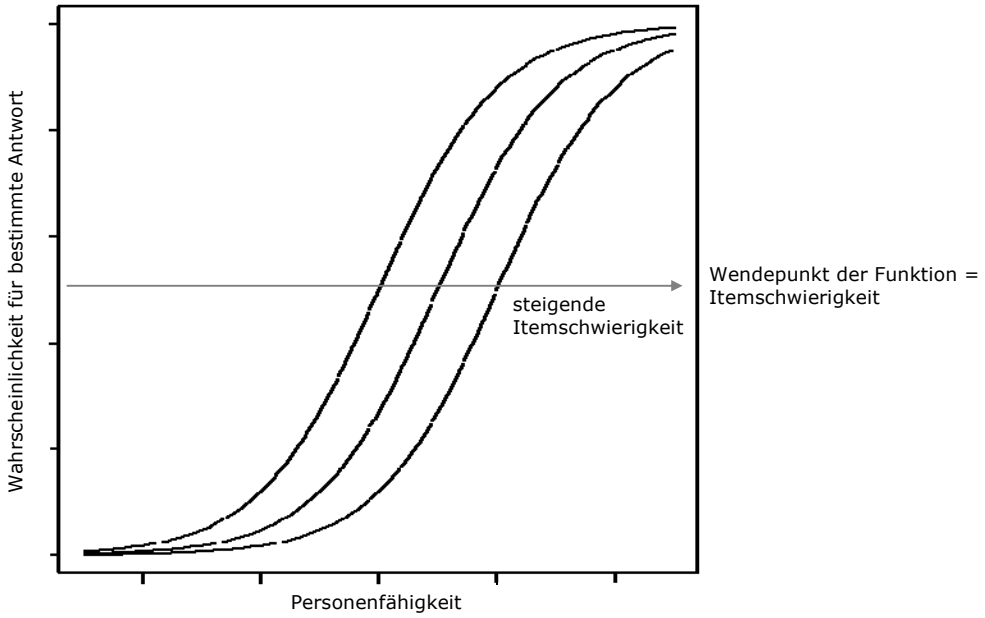


Abb. 20: *Item Characteristic Curve* (*ICC*) zu drei unterschiedlich schwierigen Items (verändert nach Bond & Fox, 2007).

Eine grundlegende Annahme des Rasch-Modells ist, dass die Bearbeitung eines Items von der Bearbeitung aller anderen Items unabhängig ist und Zusammenhänge zwischen den Items und den Personen somit vollständig durch die latente Variable, in diesem Fall Modellkompetenz, erklärt werden. Diese Annahme wird als lokale stochastische Unabhängigkeit der Items bezeichnet (Embretson & Reise, 2000).

Aus der Wahrscheinlichkeit für Antworten auf die einzelnen Items ergibt sich eine Wahrscheinlichkeit für Muster in der gesamten Datenmatrix. Die Wahrscheinlichkeit der beobachteten Datenmatrix unter den Annahmen, die im psychometrischen Modell getroffen werden, wird als *Likelihood* bezeichnet und kann Werte zwischen 0 und 1 annehmen (Moosbrugger, 2012; Rost, 2004). Sie ergibt sich aus den Wahrscheinlichkeiten der Antwortmuster aller Personen und wird in der Regel als logarithmierte *log-Likelihood* berechnet, die von -∞ bis 0 geht. Hierfür wird in einem iterativen Schätzverfahren, der *Maximum Likelihood Estimation*, die Passung zwischen den empirischen Daten und dem theoretischen Modell maximiert. Die geschätzten Werte für die Modellparameter konvergieren dabei gegen die optimalen Werte (Rost, 2004). Die Passung der Werte wird als *Deviance* angegeben, die den Unterschied zwischen der *log-Likelihood* des iterativ geschätzten und eines perfekt passenden Modells angibt. Sie wird als negative doppelte *log-Likelihood* des geschätzen Modells berechnet (Senkbeil et al., 2005).

Aus unidimensionalen Modellen wie dem Rasch-Modell leiten sich multidimensionale IRT-Modelle (MIRT-Modelle) ab. Sie erlauben differenzierte Aussagen über Kompetenzen und Teilkompetenzen von Personen und den Vergleich konkurrierender Modelle (Reckase, 1997). Statt in einer Fähigkeitsdimension wird in MIRT-Modellen die Personenfähigkeit als Funktion mehrerer Teilkompetenzen kompensatorisch modelliert, wenn nicht für jede Teilkompetenz Personenfähigkeiten geschätzt werden (Hartig, 2008; Reckase, 1997).

Im vorliegenden Projekt wurde für die Schätzung ein- sowie mehrdimensionaler Modelle die Software ConQuest (Version 2.0; Wu et al., 2007) verwendet. Diese nutzt die *Marginal Maximum Likelihood Estimation* mit einem *Expectation-Maximization*-(EM-)Algorithmus als Schätzmethode. Hierfür wird Verteilung der Personenfähigkeiten in Segmente unterteilt. Jedem Segment wird jeweils ein repräsentativer Wert (*node*) und eine Auftretenswahrscheinlichkeit (*weight*) zugewiesen (Embretson & Reise, 2000). In einem ersten Schritt, dem *Expectation*-Schritt, wird berechnet, welche Anzahl an Personen auf jedem *trait level* (repräsentiert durch die *nodes*) erwartet wird und welche Anzahl an Personen die einzelnen Items lösen sollte. Diese Erwartungen gehen in

den zweiten Schritt, den *Maximization*-Schritt, ein und werden zur Schätzung von Itemparametern verwendet. Diese werden im nächsten *Expectation*-Schritt wiederum zur Schätzung der erwarteten Personenanzahlen eingesetzt. Wenn sich die Parameter nur noch gering verändern, wird die Schätzung abgebrochen (Embretson & Reise, 2000; Rost, 2004).

Für ein- bis dreidimensionale Modelle empfehlen Wu et al. (2007) als Schätzalgorithmus die *Gauss-Hermite Quadrature* nach Bock und Aitken (1981); für Modelle ab vier Dimensionen die Monte Carlo-Methode nach Volodin und Adams (1995). Diese unterscheiden sich darin, wie die *nodes* und deren Auftretenswahrscheinlichkeit eingebunden werden (Wu et al., 2007). Mit diesen Verfahren wird zum einen ein *Item Response*-Modell, das *multidimensional random coefficients multinomial logit model* (Adams, Wilson & Wang, 1997), geschätzt. Es beschreibt den Prozess der Itembearbeitung in Abhängigkeit von der (latenten) Personenfähigkeit (Wu et al., 2007). Zum anderen wird ein Populationsmodell, und zwar ein multivariates Regressionsmodell, geschätzt. Dieses erlaubt die Unterscheidung von Subpopulationen, die sich in den Prädiktoren wie z. B. Alter oder allgemeinen kognitven Fähigkeiten unterscheiden. Je nach Ausprägung dieser Variablen wird für eine Person eine spezifische Wahrscheinlichkeitsverteilung der Personenfähigkeit angenommen. Diese Verteilung, auch a posteriori-Verteilung genannt, beruht auf latenten Regressionen, die Zusammenhänge von Prädiktoren mit der Personenfähigkeit beschreiben. Sie kann dadurch zu einer genaueren Vorhersage der Personenfähigkeiten anhand dieser weiteren Personenmerkmale genutzt werden (Wu et al., 2007).

4.2.2 Behandlung fehlender Werte

Im Anschluss an die Dateneingabe wurden 20 % der Daten auf Eingabefehler überprüft. 0.03 % der Daten waren fehlerhaft, so dass davon ausgegangen werden kann, dass der Fehleranteil im gesamten Datensatz keine nennenswerte Rolle spielt. 42 Schülerinnen und Schüler (3.57 %) hatten bei mehr als der Hälfte der MC-Items doppelt oder nichts angekreuzt. Sie wurden von der Datenauswertung ausgeschlossen, weil nicht davon ausgegangen werden kann, dass sie ausreichend viele Items konstruktgemäß bearbeitet haben. Die

Stichprobengröße reduziert sich damit für alle Analysen auf $N = 1136$ ($n_{Jgst.\ 7} = 304$; $n_{Jgst.\ 8} = 286$; $n_{Jgst.\ 9} = 273$; $n_{Jgst.\ 10} = 273$).

Jede Schülerin und jeder Schüler bearbeitete mit neun Items ein Fünftel des Itempools (Kapitel 4.1)[12]. Im Datensatz fehlten deshalb designbedingt ca. 80 % der Daten zu den Modellkompetenz-Items. Tab. 13 stellt die Anteile fehlender Werte in den übrigen Variablen dar.

Tab. 13: Prozentuale Anteile fehlender Werte im Datensatz.

Variable	Alter	Geschlecht	allgemeine kognitive Fähigkeiten	Lese-geschwindig-keit	Leseverstehen	NOS
fehlende Werte [%]	0.79	0.79	6.95	0	0	16.02

Fortsetzung Tab. 13.

Variable	Biologie-note	Chemienote	Physiknote	Mathe-matiknote	Deutsch-note	Note erste Fremd-sprache
fehlende Werte [%]	4.05	10.65	4.23	1.85	2.11	2.82

Die Chemienote fehlte mit 10.65 % relativ häufig. Wenn man sich den Datenausfall auf dieser Variable genauer anschaut, stellt man fest, dass 106 von 121 fehlenden Werten (87.60 %) in der Jahrgangsstufe 7 auftraten. Hier ist zu vermuten, dass diese Schülerinnen und Schüler noch keinen Chemieunterricht hatten, so dass für sie keine Chemienote angegeben werden kann. Diese Variable wurde deshalb in Dummy-Variablen umcodiert, die eine zusätzliche Kategorie für diese Schülerinnen und Schüler beinhalteten. Wie mit dieser

[12] Für die Anzahl an Bearbeitungen je Item siehe Anhang 9.

Variable in den darauf aufbauenden Analysen verfahren wurde, ist in Kapitel 4.2.6 beschrieben.

Der Umgang mit fehlenden Werten wird schon lange als wichtiger Punkt der Datenauswertung thematisiert (Rubin, 1976). Am häufigsten wird zur Behandlung fehlender Werte der fall- oder paarweise Ausschluss verwendet, d. h. es werden nur die Personen in die Analysen einbezogen, die keine fehlenden Werte aufweisen (Peugh & Enders, 2004). Dies ist in der Regel mit den Problemen einer stark verzerrten Parameterschätzung und einer ggf. stark reduzierten Stichprobengröße verbunden (Allison, 2002; Schafer & Graham, 2002; Little & Rubin, 2002). Dieses Vorgehen wird deshalb nur dann als akzeptabel eingeschätzt, wenn ein geringer Anteil der Stichprobe (unter 5 %) ausgeschlossen wird (Enders, 2010; Graham, Cumsille & Elek-Fisk, 2003; Schafer, 1997; Graham, 2009; Lüdtke, Robitzsch, Trautwein & Köller, 2007). Während dieses Vorgehen demnach für die 3.57 % der Schülerinnen und Schüler, die nur einen geringen Anteil der Modellkompetenz-Items bearbeitet haben, zulässig ist, würde es für die weiteren Variablen im hier verwendeten Datensatz mit 291 Schülerinnen und Schülern 25.8 % des Datensatzes betreffen. Außerdem setzen viele statistische Standardverfahren vollständige Datensätze voraus (Little & Rubin, 2002).

Zur Auswahl eines Vorgehens werden Annahmen über den statistischen Zusammenhang zwischen den Daten und dem Vorliegen von fehlenden Werten getroffen. Diese beziehen sich nicht auf das kausale Zustandekommen der fehlenden Werte und dürfen entsprechend nicht als Aussagen über die *Gründe* für den Datenausfall verstanden werden (Enders, 2010; Graham, 2009; Lüdtke & Robitzsch, 2010). Rubin (1976) nimmt in seiner Theorie zu Ausfallprozessen an, dass das Fehlen einer Variable wahrscheinlichkeitsverteilt ist, und führt für diese Verteilung die Indikatorvariable R ein. Diese gibt an, ob ein Wert auf einer spezifischen Variable fehlt, und kann mit Variablen im Datensatz in Verbindung stehen. Auf dieser Grundlage klassifiziert Rubin (1976) drei Mechanismen, wie die Wahrscheinlichkeit eines fehlenden Werts mit den Daten zusammenhängt (Abb. 21). Die Klassfikation des Ausfallprozesses hängt

dabei immer vom jeweiligen Datensatz und der jeweiligen Fragestellung ab (Graham, 2009; Schafer, 1997)[13].

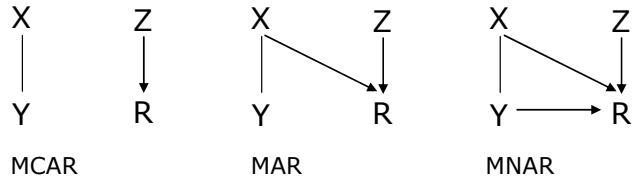


Abb. 21: Grafische Darstellung von Missing Data-Mechanismen (Schafer & Graham, 2002). X und Y sind Variablen im Datensatz, die miteinander zusammenhängen, Z eine Variable für Ursachen für Antworten auf der Variable Y, R eine Indikatorvariable, die angibt, ob in Y eine Antwort vorliegt.

Die Mechanismen sind folgendermaßen definiert (Rubin, 1976):

1. *Missing Completely At Random* (MCAR): Werte fehlen rein zufällig, z. B. bei einem Multi-Matrix-Design (Kapitel 4.1), auch als *planned missingness* (Graham, 2009, S. 565) oder *missing by design* bezeichnet (Schafer, 1997). Die Wahrscheinlichkeit eines fehlenden Werts hängt demnach weder mit beobachteten noch unbeobachteten Daten zusammen.

2. *Missing At Random* (MAR): Die Wahrscheinlichkeit eines fehlenden Werts auf einer bestimmten Variablen steht in systematischer Beziehung zu einer oder mehreren beobachteten Variablen, aber nicht mit der (unbeobachteten) Ausprägung der Variablen selbst (deshalb z. T. auch *Conditionally Missing At Random* bezeichnet, Graham et al., 2003, S. 89; Graham, 2009, S. 553).

---

[13] Für Beispiele hierzu siehe Graham (2009).

3. *Missing Not At Random* (MNAR): Auch nach der Kontrolle der beobachteten Variablen hängt das Auftreten der Werte von der Ausprägung der fehlenden Werte selbst ab. Er kann deshalb nicht sinnvoll geschätzt werden.

Verfahren wie die Schätzung von Rasch- oder MIRT-Modellen, die auf *Maximum Likelihood Estimation* beruhen, benötigen keine Information über den Ausfallprozess, wenn MAR oder MCAR gilt und kein Populationsmodell (Kapitel 4.2.1) einbezogen werden soll (Little & Rubin, 2002). Dies trifft z. B. auf die Modellkompetenz-Items zu, die den Schülerinnen und Schülern in einem BIBD vorgegeben wurden (vgl. Kapitel 4.1). Da die angestrebten Analysen zur Strukturierung (Kapitel 4.2.3), Graduierung (Kapitel 4.2.4) und Beschreibung von Modellkompetenz im Querschnitt (Kapitel 4.2.5) keine vollständigen Daten auf diesen Variablen voraussetzen, können die designbedingt fehlenden Daten ignoriert werden. Modellkompetenz-Items, die vorgegeben, aber nicht beantwortet wurden, wurden für die Schätzung sowohl der Itemparameter als auch der Personenfähigkeiten als falsch gewertet. Da die Schülerinnen und Schüler das nächste Testheft erst erhielten, wenn sie angaben, alle Aufgaben bearbeitet und kontrolliert zu haben, konnten die fehlenden Werte nicht aus Zeitmangel zustande kommen. Schülerinnen und Schüler, die stark lückenhaft antworteten, wurden aus dem Datensatz ausgeschlossen (s. o.). Deshalb können die fehlenden Werte hier als mangelnde Kompetenz interpretiert werden.

Als Grundlage für die Berechnung von Korrelationen, die Beziehungen von Modellkompetenz zu anderen Konstrukten abbilden, wird ein vollständiger Datensatz für die relevanten Variablen benötigt (Kapitel 4.2.6). Aus diesem Grund wurde als Methode zum Umgang mit fehlenden Werten die multiple Imputation gewählt, die neben Verfahren der *Maximum Likelihood Estimation* hierfür empfohlen wird (Graham, 2009; Graham et al., 2003; Little & Rubin,

2002; Lüdtke et al., 2007; Peugh & Enders, 2004; Schafer & Graham, 2002)[14]. Diese beiden Verfahren bieten im Vergleich zu klassischen den Vorteil, dass sie Simulationsstudien zufolge zu valideren Ergebnissen führen, selbst wenn MAR bzw. MCAR fälschlicherweise angenommen wird (Collins, Schafer & Kam, 2001).

Bei der Multiplen Imputation wird jeder fehlende Wert durch mehrere sinnvolle mögliche Werte ersetzt. Auf diese Weise kann die Unsicherheit der Imputation durch Unterschiede zwischen diesen Werten abgebildet werden, es müssen keine Personen aus den Analysen ausgeschlossen werden und es kann mit mehreren vollständigen Datensätzen weitergearbeitet werden (Lüdtke et al., 2007, Graham, 2009). Ziel ist dabei nicht, die individuellen Werte zu schätzen, sondern wichtige Charakteristika des gesamten Datensatzes in Bezug auf die Population zu erhalten, z. B. Mittelwerte und Varianzen (Graham, 2009).

Die Imputation der Werte ist der erste von drei Schritten, die nach diesem Verfahren durchgeführt werden müssen (Rubin, 1987). Hierfür wird ein Imputationsmodell mit den Variablen des Analysemodells sowie ggf. zusätzlichen Hilfsvariablen spezifiziert (*inclusive strategy*; Collins et al., 2001; Graham, 2009), die mit dem Ausfall korrelieren und ggf. MNAR in Richtung MAR reduzieren können (Collins et al., 2001). Beim Ansatz der *Fully conditional specification* (*FCS*)[15] werden die Daten ausgehend von zufälligen Werten Variable für Variable imputiert (van Buuren, Brand, Groothuis-Oudshoorn & Rubin, 2006; van Buuren, 2007). Hierfür wird für jede Variable Y ein spezifisches univaria-

---

[14] Für eine Diskussion von Vor- und Nachteilen klassischer, imputationsbasierter und modellbasierter Verfahren zur Behandlung fehlender Werte siehe z. B. Graham (2009), Little und Rubin (2002) sowie Lüdtke et al. (2007).

[15] Van Buuren (2007) nennt als weitere verwendete Bezeichnungen für FCS *stochastic relaxation*, *variable-by-variable imputation*, *regression switching*, *sequential regressions*, *ordered pseudo-Gibbs sampler*, *partially incompatible MCMC*, *iterated univariate imputation* und *chained equations*.

tes Imputationsmodell formuliert, das die Wahrscheinlichkeit von Werten für diese Variable in Abhängigkeit von beobachteten Werten für andere relevante Variablen (z. B. Hilfsvariablen) spezifiziert. Daraus resultiert eine bedingte Wahrscheinlichkeitsverteilung für Werte, die Y annimmt (van Buuren et al., 2006). Meistens reicht es aus, die 15 erklärungsstärksten Variablen im Datensatz ins Imputationsmodell einzubeziehen. Dabei ist es zulässig, Variablen mit fehlenden Werten als Prädiktoren zu verwenden (van Buuren & Groothuis-Oudshoorn, 2011).

Aus dieser Verteilung wird für die entsprechende Variable z. B. durch eine lineare oder logistische Regression mithilfe der beobachteten Fälle auf der Variable und mithilfe der Prädiktoren ein Wert gezogen, mit dem der fehlende ersetzt wird. Eine Iteration entspricht einem Durchlauf über alle Imputationsmodelle für die Variablen mit fehlenden Werten[16]. Für die erste Variable werden die fehlenden Werte zufällig gesetzt. Für die weiteren Variablen werden die fehlenden Werte auf Grundlage der Werte imputiert, mit denen in den vorherigen Schritten die anderen Variablen vervollständigt wurden (van Buuren et al., 2006; van Buuren & Oudshoorn, 1999; van Buuren, 2007). Bei diesem Verfahren reichen bereits fünf bis zwanzig Iterationen aus, um zuverlässig adäquate Werte zu imputieren (van Buuren et al., 2006; van Buuren, 2007). Im Vergleich zu anderen Ansätzen wie dem *Joint Modelling*, das für alle Variablen ein gemeinsames Imputationsmodell mit einer gemeinsamen Imputationsmethode erfordert, hat die *FCS* den Vorteil einer höheren Flexibilität, da für jede Variable sowohl ein spezifisches Imputationsmodell spezifiziert als auch eine spezifische Imputationsmethode ausgewählt werden kann (van Buuren et al., 2006; van Buuren, 2007).

---

[16] Die Ziehungen von Werten auf der Grundlage der Imputationsmodelle wird auch als *Gibbs sampling* bezeichnet (van Buuren & Oudshoorn, 1999). Es zielt darauf ab, die unbekannte gemeinsame Wahrscheinlichkeitsverteilung zu approximieren, die als Grundlage der einzelnen univariaten Imputationsmodelle angenommen wird.

Im hier vorgestellten Projekt wurden zwanzig Iterationen durchgeführt, bevor ein Datensatz abgespeichert wurde. Insgesamt wurden zehn vollständige Datensätze erstellt (vgl. Graham, 2009). Dafür wurde das Verfahren der *FCS* eingesetzt, das in *Multivariate Imputation by Chained Equations* (*MICE*) als R-Paket implementiert ist (van Buuren & Groothuis-Oudshoorn, 2011). *MICE* ist für nominale und ordinale Daten wie z. B. Geschlecht und Schulnoten angemessener als z. B. das häufig eingesetzte Paket NORM, das nach dem *Joint Modelling*-Ansatz arbeitet (Lüdtke et al., 2007). Wenn für weitere Analysen lineare Zusammenhänge zwischen den Variablen angenommen werden, können die Daten mit der Imputationsmethode *Predictive Mean Matching* (*pmm*) imputiert werden (van Buuren & Groothuis-Oudshoorn, 2011). In der hier verwendeten Version 2.10 (R-Version 2.14.0) sowie den Funktionen micetrun (micetrun_0.6-17.R) erstellt *MICE* automatisch eine Prädiktormatrix, die relevante Variablen für die Imputation angibt. Diese werden auf der Grundlage von Korrelationen zwischen dem Auftreten von fehlenden Werten auf einer Variable (R) mit anderen Variablen, Korrelationen zwischen Variablen im Datensatz und dem Anteil an beobachteten Fällen je Variable ausgewählt (van Buuren & Groothuis-Oudshoorn, 2011). Simulationen zeigen, dass daraus eine verbesserte Parameterschätzung mit einer größeren Effizienz und einem geringeren Bias resultiert (van Buuren et al., 2006; Collins et al., 2001).

Inwiefern die Imputationen adäquat sind, kann zum einen darüber geprüft werden, ob sie beobachtbar gewesen wären. Für Schulnoten den Wert 7 zu imputieren, wäre z. B. nicht angemessen. Darüber hinaus kann das Konvergenzverhalten der Imputationen gegen die definierten Verteilungen grafisch geprüft werden, indem die imputierten Werte gegen die Iterationszahl abgebildet werden (Peugh & Enders, 2004). Dabei sollten über die Iterationen kein Trend, sondern möglichst unsystematische Muster erkennbar sein, deren Kurven sich überschneiden (van Buuren & Oudshoorn, 1999). Ein sichtbarer Trend würde bedeuten, dass die aufeinander folgenden Iterationen nicht in ausreichendem Ausmaß voneinander unabhängig sind (Peugh & Enders, 2004).

Der zweite Schritt nach der Imputation der Daten ist die *Durchführung der angestrebten Analysen* mit den vollständigen Datensätzen (Kapitel 4.2.6). Die so erhaltenen Ergebnisse werden in einem dritten Schritt, dem *Pooling*, zusammengefasst. Da die fehlenden Werte jeweils durch mehrere Werte ersetzt wurden, führen die Analysen der vollständigen Datensätze zu voneinander abweichenden Ergebnissen, die die Unsicherheit bei der Imputation abbilden. Die Ergebnisse der Analysen werden zur Zusammenfassung gemittelt. Die Standardabweichung wird als Summe der *within*- und *between*-Varianz der Imputationen berechnet (Rubin, 1987; Peugh & Enders, 2004).

4.2.3 Überprüfung der Strukturierung von Modellkompetenz

Die MC-Items zur Erfassung von Modellkompetenz wurden so entwickelt, dass sie eine Kombination aus einer spezifischen Teilkompetenz und einem Niveau abbilden. Es wird deshalb angenommen, dass zur Lösung eines Items jeweils ausschließlich diese Teilkompetenz benötigt wird. Die verwendeten MIRT-Messmodelle bestehen demnach aus mehreren unidimensionalen Skalen (*between-item multidimensionality*; Adams et al., 1997).

Mit der Software ConQuest (Version 2.0; Wu et al., 2007) wurden drei hypothetische Strukturmodelle (ein-, zwei- bzw. fünfdimensional; vgl. Kapitel 2.5) unabhängig voneinander geschätzt. Außerdem wurden als Kontrollmodelle zwei Dummy-Modelle gerechnet. Diese bestanden in einem zwei- und einem fünfdimensionalen Modell, denen zufällig Items zugewiesen wurden. Damit konnte direkt die Passung eines theoriegeleiteten und eines zufällig zugeordneten Modells verglichen werden. Das ein- und die zweidimensionalen Modelle wurden, der Empfehlung von Wu et al. (2007) folgend, mit dem Gauss-Hermite Quadrature-Verfahren geschätzt, die fünfdimensionalen mit der Monte Carlo-Methode. Dabei wurde jeweils zunächst eine grobe Schätzung mit sieben *nodes* je Dimension durchgeführt, um einen sinnvollen Bereich für die Parameter einzugrenzen. Die so berechneten Parameter wurden als Startwerte für eine genauere Schätzung mit 20 *nodes* je Dimension eingelesen. Dieses Vorgehen führt zu einer effizienten, genauen Schätzung. Item A2.3 wurde mit einem grenzwertigen *Itemfit* (*wMNSQ* = 1.08; t = 2.0) im Itempool für diese Skalierungen belassen.

Die *Deviance* auf diese Weise geschätzter Modelle ist χ^2-verteilt. Wenn Modelle durch Restringierungen aufeinander reduziert werden können, d. h. wenn ein komplexes Modell in einem einfacheren enthalten ist, kann die *Deviance* dieser Modelle direkt über einen χ^2-Differenzentest verglichen werden (Senkbeil et al., 2005). Dabei entspricht die Differenz der Parameter, die für die Modelle berechnet werden, den Freiheitsgraden der Verteilung.

Eine zweite Möglichkeit, die geschätzten Strukturmodelle miteinander zu vergleichen, bieten Informationskriterien: *Akaike Information Criterion* (*AIC*), *Bayesian Information Criterion* (*BIC*), *consistent AIC* (*cAIC*) sowie *sample size adjusted BIC* (ssBIC). Diese beziehen ebenfalls die *Likelihood* der Modelle sowie die Zahl der Parameter, die für die Modelle berechnet werden, ein (Rost, 2004).

Eine dritte Möglichkeit, die empirische Dimensionalität von Konstrukten zu prüfen, bieten latente Korrelationen zwischen den Dimensionen entsprechender Strukturmodelle. Diese Korrelationen werden von ConQuest im Rahmen der Modellschätzung ebenfalls ausgegeben. Hohe Korrelationen sprechen für ein eindimensionales Konstrukt, niedrige für mehrere Dimensionen. Dabei ist zu beachten, dass latente Korrelationen im Gegensatz zu manifesten messfehlerbereinigt sind und deshalb in der Regel höher als diese ausfallen (Carstensen et al., 2007).

4.2.4 Überprüfung der Graduierung von Modellkompetenz

Die Definition von Teilkompetenzen und Niveaus ermöglicht die hypothetische Beschreibung von schwierigkeitsbestimmenden Itemmerkmalen. Diese wurden auf der Grundlage der Theorie zu Modellen und Modellkompetenz (Kapitel 2.4.1) mit Blick auf Prozesse bei der Aufgabenbearbeitung und situative Merkmale formuliert. Ziel der Itementwicklung im hier vorgestellten Projekt war, die Schwierigkeit der Items durch die Zuordnung zu den a priori formulierten, hypothetischen Niveaus und daraus abgeleiteten Itemmerkmalen zu variieren (vgl. Kapitel 3.1.3, 3.1.5). Weitere schwierigkeitsrelevante Itemmerkmale, die nicht in Verbindung zur Theorie stehen, z. B. Einbindung von Abbildungen, Fachwissen etc. (Kapitel 3.1.1), wurden in den Aufgaben stan-

dardisiert. Indem geprüft wird, inwiefern die hypothetischen Niveaus sich empirisch abbilden lassen, können sowohl Rückschlüsse auf die Repräsentation des Kompetenzmodells durch den Test als auch auf die Struktur von Modellkompetenz gezogen werden (Borsboom et al., 2004; Embretson, 1983; Hartig & Jude, 2007).

Die Itemschwierigkeit kann klassisch als Lösungshäufigkeit oder im Rahmen der der IRT-Modellierung als Itemparameter berechnet werden. Die Lösungshäufigkeiten der MC-Items basieren aufgrund des Multi-Matrix-Designs (Kapitel 4.1) jedoch nicht auf jeweils denselben Schülerinnen und Schülern. Als Kennwerte für die Schwierigkeit der Items wurden deshalb die Itemschwierigkeiten aus der Skalierung des Strukturmodells verwendet, das die beste Passung mit den Daten aufwies (Kapitel 4.3.1, 4.4.1). Dabei gingen nur die Items in die Berechnung ein, die nach psychometrischen Kriterien selektiert wurden (Kapitel 3.1.6) und bei deren Bearbeitung die kognitiven Prozesse der Schülerinnen und Schüler sich auf den jeweiligen Teil des Kompetenzmodells beziehen ließen (Kapitel 3.3).

Zur empirischen Überprüfung der Graduierung von Modellkompetenz in Niveaustufen (Kapitel 2.4.1) wurde eine Varianzanalyse durchgeführt. Dieses Verfahren setzt normalverteilte Werte und Varianzhomogenität voraus, so dass ein erster Schritt diese Voraussetzungen durch den Kolmogorov-Smirnov-Test bzw. den Levene-Test prüft (Zöfel, 2003). Im Anschluss wurden die Mittelwerte der Itemschwierigkeiten je Niveau berechnet. Da die Niveaus a priori als ordinal beschrieben werden (Kapitel 2.4.1), wurden Unterschiede zwischen den einzelnen Niveaus nicht durch post hoc-Tests verglichen, sondern es wurde ausschließlich geprüft, ob die Itemschwierigkeit mit den Niveaus im Mittel ansteigen und ob diese Unterschiede signifikant sind. Dies wurde inferenzstatistisch durch eine einfaktorielle Varianzanalyse getestet, die anhand einer *F*-Verteilung überprüft, ob die Varianz zwischen den Niveaus größer als innerhalb der Niveaus ist (Backhaus, Erichson, Plinke & Weiber, 2006). Als abhängige Variable ging hierfür die Itemschwierigkeit ein, als unabhängige die Niveauzugehörigkeit der Items mit den Niveaus als drei Faktorstufen. Der

Quotient aus der Varianz zwischen den Niveaus und der innerhalb der Niveaus ist die Effektstärke η^2 (Sedlmeier & Renkewitz, 2008).

4.2.5 Beschreibung von Modellkompetenz im Querschnitt

Für die Beschreibung von Modellkompetenz im Querschnitt wurden die Personenfähigkeiten der Schülerinnen und Schüler betrachtet. Wenn Aussagen auf Populationsebene und nicht auf der Ebene individueller Schülerinnen und Schüler getroffen werden sollen, sind *plausible values* (PVs) als Schätzwerte für Personenfähigkeiten am besten geeignet. Sie geben Charakteristika der Population (z. B. Mittelwerte oder Varianzen) unverzerrt wieder und können als multiple Imputationen für die Personenfähigkeit verstanden werden, deren Unterschiede wie die anderer imputierter Werte die Unsicherheit der Schätzung abbildet (Wu, 2005).

Wenn PVs im Kontext von Analysen verwendet werden, in denen die Personenfähigkeiten die abhängige Variable darstellen, müssen die unabhängigen Variablen als Hintergrundvariablen in ihre Berechnung eingehen. Dies zielt darauf ab, Personenschätzer zu erhalten, in denen der Anteil dieser Variablen an der geschätzten Personenfähigkeit auspartialisiert wurde. Wenn unabhängige Variablen kategorial sind, müssen sie in Dummy-Variablen umcodiert werden (Wu, 2005). Für die Schätzung von Modellkompetenz-PVs wurde analog zum Vorgehen bei der Strukturierung von Modellkompetenz (Kapitel 4.2.3) zunächst ohne Hintergrundmodell ein Modell mit sieben *nodes* geschätzt, dessen Parameter als Startwerte in die Schätzung einer latenten Regression mit 20 *nodes* und der Jahrgangsstufe als unabhängiger Variable eingingen. Hierfür wurde die Jahrgangsstufe als Dummy-Variable mit Jahrgangsstufe 7 als Referenz codiert, d. h. es wurde jeweils eine Variable für Jahrgangsstufe 8 bis 10 eingefügt, die den Wert 1 annimmt, wenn die Person in dieser Jahrgangsstufe ist, und 0, wenn nicht.

Deskriptiv können Modellkompetenz-Unterschiede zwischen den Jahrgangsstufen über die Mittelwerte und Standardfehler der PV*s* beschrieben werden. Die Standardfehler der PVs werden analog zu imputierten Daten berechnet (Wu, 2005). Die Ausgabe der latenten Regression in ConQuest umfasst un-

standardisierte latente Regressionsgewichte, die als Effektgrößen der Prädiktoren interpretiert werden können, sowie deren Standardfehler. Die Regressionsgewichte stellen bei der hier verwendeten Dummy-Codierung die Mittelwertsdifferenzen im Vergleich zu Jahrgangsstufe 7 dar (Wu et al., 2007). Als Maß für die Varianzaufklärung durch die Variablen Jahrgangsstufe und Alter wurde das Bestimmtheitsmaß R^2 aus der Varianz mit sowie der Varianz ohne latente Regression berechnet. Dies ermöglichte außerdem über den Quotienten aus der Varianz zwischen den Niveaus und der Varianz innerhalb der Niveaus einen F-Test und somit eine Signifikanzprüfung der gesamten Regressionsanalyse. Die einzelnen Regressionskoeffizienten wurden mit t-Statistiken auf ihre Signifikanz geprüft, indem die Regressionsgewichte durch ihre Standardfehler geteilt wurden (Backhaus et al., 2006).

4.2.6 Beziehungen von Modellkompetenz zu anderen Konstrukten

Um die Erlernbarkeit und Domänenspezifität von Modellkompetenz (Kapitel 2.1.1) sowie ihren Bezug zum Wissenschaftsverständnis (Kapitel 2.4.2) empirisch zu untersuchen, wurden die Beziehungen von Modellkompetenz zum Wissenschaftsverständnis, allgemeinen kognitiven Fähigkeiten, Lesefähigkeiten, Geschlecht sowie Schulnoten als Indikatoren für Leistungen in sprachlichen sowie naturwissenschaftlichen Fächern betrachtet. Diese Zusammenhänge können im Sinne einer konvergenten und diskriminanten Validierung von Modellkompetenz interpretiert werden.

Analog zur Beschreibung von Modellkompetenz im Querschnitt (Kapitel 4.2.5) wurden als Personenschätzer PVs ausgeschrieben. Hierfür wurden die Variablen einbezogen, deren Beziehung zu Modellkompetenz untersucht werden soll und die deshalb über die multiple Imputation vervollständigt wurden (Wissenschaftsverständnis, allgemeine kognitive Fähigkeiten, Lesefähigkeiten, Geschlecht sowie Schulnoten). Für die Chemienote wurden keine Werte imputiert, da anzunehmen ist, dass der Großteil der fehlenden Werte darauf zurückgeführt werden kann, dass die jeweiligen Schülerinnen und Schüler noch keinen Chemieunterricht hatten (Kapitel 4.2.2). Es wurde deshalb eine zusätzliche Variable eingeführt, die angibt, ob die Person bislang Chemieunterricht hatte. Diese ist für den gesamten Datensatz vollständig und ging des-

halb gemeinsam mit den anderen Variablen in das Hintergrundmodell zur Schätzung der PVs ein.

Bei der Skalierung der zehn imputierten Datensätze mit Hintergrundmodell wiesen Item E2.4 (M_{wMNSQ} = 1.15; M_t = 3.1) sowie Item A2.3 (M_{wMNSQ} = 1.12; M_t = 2.7) einen schlechten *Itemfit* auf. Sie wurden deshalb aus den Analysen ausgeschlossen. Im verbleibenden Itempool war der t-Wert von Item Z2.6 mit t = 2.0 (*wMNSQ* = 1.10) in einem Datensatz grenzwertig. Da es in allen anderen Datensätzen jedoch einen akzeptablen *Itemfit* aufwies, wurde es im Itempool für diese Berechnungen belassen.

Für die Berechnung der Korrelationen zwischen den PVs und der Chemienote wurde der Datensatz mit einem fallweisen Ausschluss auf n = 1015 reduziert. Auch wenn hierfür 10.65 % der Fälle – und damit mehr als 5 % wie in der Literatur empfohlen (Enders, 2010; Graham, Cumsille & Elek-Fisk, 2003; Schafer, 1997; Graham, 2009; Lüdtke, Robitzsch, Trautwein & Köller, 2007) – ausgeschlossen wurden, erscheint dieses Vorgehen am angemessensten. Die Chemienote könnte nur für alle Schülerinnen und Schüler gleichzeitig vervollständigt werden; dies würde dann aber auch die umfassen, die vermutlich noch nicht in Chemie unterrichtet wurden. Für die Imputation dieser Variable könnte demnach nur mit einem Teildatensatz gearbeitet werden, der dennoch 9.36 % der Schülerinnen und Schüler ausschließen müsste. Auch wenn die Schätzung der PVs vermutlich etwas genauer wäre, rechtfertigt der Unterschied somit aus pragmatischer Sicht nicht die zusätzliche Imputation für die 15 Schülerinnen und Schüler in den Jahrgangsstufen 8 bis 10. Für die Korrelation von Modellkompetenz mit der Chemienote wurde deshalb ein Hintergrundmodell verwendet, das die gleichen Hintergrundvariablen wie für die übrigen Korrelationen sowie zusätzlich die Chemienote beinhaltete. Die Variable Chemieunterricht wurde hier aus dem Hintergrundmodell ausgeschlossen, da durch den Ausschluss der Fälle mit fehlenden Werten alle Schülerinnen und Schüler die gleiche Ausprägung für diese Variablen aufwiesen.

In der Skalierung für die Chemienote passten Item E2.4 (M_{wMNSQ} = 1.13; M_t = 2.6), Item E3.4 (M_{wMNSQ} = 0.87; M_t = -2.5) sowie in neun von zehn Da-

tensätzen Item A2.3 (M_{wMNSQ} = 1.09; M_t = 2.1) nicht adäquat zum Messmodell; in einem Datensatz wies Item Z2.6 einen grenzwertigen t-Wert auf (t = 2.0; *wMNSQ* = 1.10). Diese Items wurden deshalb aus den Berechnungen ausgeschlossen. Z2.6 verschlechtert sich in einer dritten Skalierung in allen Datensätzen (M_{wMNSQ} = 1.11; M_t = 2.1) und wird deshalb für diese Skalierung ebenfalls ausgeschlossen.

Die Beziehungen von Modellkompetenz zu anderen Konstrukten werden in den einzelnen imputierten Datensätzen in IBM SPSS Statistics (Version 19) als Produkt-Moment-Korrelationen[17] berechnet (Bortz & Döring, 2005). Da die Variablen, die korreliert werden, ins Hintergrundmodell einbezogen bzw. mit dem Hintergrundmodell berechnet wurden, sind die Korrelationen messfehlerbereinigt (Rost, 2004). Die Ergebnisse dieser Analysen werden als arithmetische Mittel zusammengefasst. Die Variabilität wird in einer Standardabweichung abgebildet, die als Summe der *within*- und *between*-Varianz der Imputationen berechnet wird (Rubin, 1987; Peugh & Enders, 2004).

[17] Für die dichotomen Variablen werden diese als punktbiseriale Korrelationen bezeichnet (Bortz & Döring, 2005).

4.3 Ergebnisse

Mit den beschriebenen Methoden wurden die Daten analysiert, um Aussagen zur Strukturierung (Kapitel 4.3.1) und Graduierung von Modellkompetenz (Kapitel 4.3.2), zur Modelkompetenz von Schülerinnen und Schülern der Jahrgangsstufen 7 bis 10 im Vergleich (Kapitel 4.3.3) sowie zu Beziehungen von Modellkompetenz zu anderen Konstrukten (Kapitel 4.3.4) treffen zu können.

4.3.1 Überprüfung der Strukturierung von Modellkompetenz

Der kleinste Wert der Informationskriterien, der die beste Passung mit den Daten anzeigt, trat beim eindimensionalen Modell auf (Tab. 14). Die Unterschiede zwischen den theoriegeleiteten Modellen in den Informationskriterien sowie zwischen den Dummy-Modellen und den entsprechenden theoriegeleiteten Modellen waren gering. Ein χ^2-Test der drei genesteten theoriegeleiteten Modelle zeigte keinen signifikanten Unterschied zwischen ihnen ($\Delta\chi^2_{1D,\ 2D}(2, 1136) = 0.886$; $p = 0.64$; $\Delta\chi^2_{1D,\ 5D}(2, 1136) = 1.211$; $p = 1.0$; $\Delta\chi^2_{2D,\ 5D}(12, 1136) = 0.325$; $p = 1.0$).

Tab. 14: Vergleich konkurrierender Strukturmodelle zu Modellkompetenz (ein-, zwei-, fünfdimensional) und zwei Dummy-Modellen (ein-, zweidimensional) als Kontrolle. $N_{Schülerinnen\ und\ Schüler} = 1136$.

Modell	*Deviance*	Parameter	BIC	AIC	CAIC	ssBIC
1D	11277.37	41	11565.82	11359.37	11606.82	56149.78
2D	11276.49	43	11579.00	11362.49	11622.00	56211.49
2D-Dummy	11279.51	43	11582.03	11365.51	11625.03	56226.19
5D	11276.16	55	11663.10	11386.16	11718.10	56606.01
5D-Dummy	11264.61	55	11651.55	11374.61	11706.55	56549.86

1D = eindimensional, 2D= zweidimensional, 5D= fünfidmensional.

Die Dimension ‚Kenntnisse über Modelle' korrelierte mit der ‚Modellbildung' auf latenter Ebene in Höhe von $r = .599$. Die Korrelationen zwischen den Teilkompetenzen im fünfdimensionalen Strukturmodell zeigt Tab. 15. Wie beim Itempool mit 45 Items (Kapitel 3.1.6) war mit dem reduzierten Itempool ($N_{Items} = 40$) die Varianz und somit auch die Reliabilität gering (Tab. 16).

Tab. 15: Latente Korrelationen zwischen den Teilkompetenzen im fünfdimensionalen Strukturmodell. $N_{Schülerinnen\ und\ Schüler}$ = 1136.

	Eigenschaften von Modellen	Alternative Modelle	Zweck von Modellen	Testen von Modellen
Eigenschaften von Modellen	_			
Alternative Modelle	.679	_		
Zweck von Modellen	.591	.566	_	
Testen von Modellen	.677	.587	.616	_
Ändern von Modellen	.682	.499	.588	.763

Tab. 16: EAP/PV-Reliabilität und Varianz für verschiedene Skalierungen in ConQuest mit dem reduzierten Itempool (N_{Items} = 40).

	Dimensionen des jeweiligen Messmodells							
	1	2		5				
		KM	MB	E	A	Z	T	Ä
Reliabilität	.448	.338	.398	.322	.381	.277	.315	.325
Varianz	.516	.697	.594	1.116	.575	.730	.576	.914

Jedes Testheft enthielt neun ItemS. KM = Kenntnisse über Modelle; MB = Modellbildung; E = Eigenschaften von Modellen; A = Alternative Modelle; Z = Zweck von Modellen; T = Testen von Modellen; Ä = Ändern von Modellen.

4.3.2 Überprüfung der Graduierung von Modellkompetenz

Insgesamt stiegen die Mittelwerte der Itemschwierigkeiten mit den Niveaus an ($M_{Niveau\ I}$ = -.8468; $SD_{Niveau\ I}$ = .9335; $M_{Niveau\ II}$ = -.4194; $SD_{Niveau\ II}$ = .6264; $M_{Niveau\ III}$ = -.0161; $SD_{Niveau\ III}$ = .6953). Die Abstände zwischen den mittleren Itemschwierigkeiten je Niveaus waren etwa gleich groß (Abb. 22).

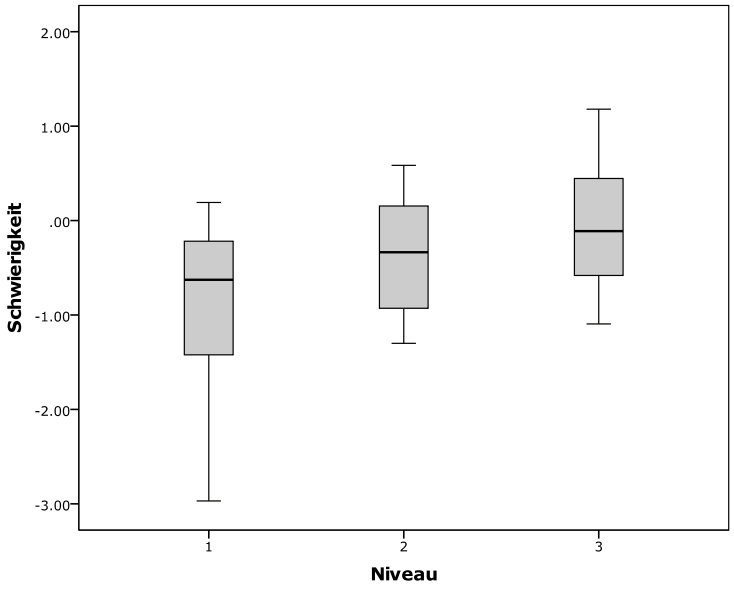


Abb. 22: Boxplots der Itemschwierigkeiten. Der Kasten zeigt den Median und die beiden mittleren Quartile an, die Linien die Extremwerte. $N_{Items\ Niveau\ I}$ = 12; $N_{Items\ Niveau\ II}$ = 14; $N_{Items\ Niveau\ III}$ = 14.

Wenn man die einzelnen Teilkompetenzen betrachtet, fällt auf, dass die Itemschwierigkeiten stark über die Niveaus streuen (Tab. 17). In der Teilkompetenz ‚Testen von Modellen' waren die meisten einfachen Items in Niveau II. Die Niveau III-Items dieser Teilkompetenz waren einfacher als die zum ‚Zweck' und ‚Ändern von Modellen'.

Tab. 17: Fallzentrierte Schwierigkeiten der einzelnen Items je Niveau und Teilkompetenz.

	Niveau I	Niveau II	Niveau III
Eigenschaften von Modellen	-.434	-0.821	-0.645
	-2.974	0.389	-0.431
Alternative Modelle	-0.694	-0.930	-0.342
	-0.176	0.155	0.137
	-1.664	0.175	0.670
Zweck von Modellen	-1.896	0.585	1.090
	-0.594	-0.769	0.336
		-0.329	-0.724
Testen von Modellen	-0.663	-1.111	-1.098
	0.173	-0.334	0.445
	0.193	-1.281	-0.583
Ändern von Modellen	-0.261	0.036	1.182
	-1.186	-0.340	-0.379
		-1.302	0.118

Da sowohl der Kolmogorov-Smirnov-Test ($D(40) = .079$, $p_{Itemschwierigkeit} = .200$) als auch der Levene-Test ($F(2,37)_{Itemschwierigkeit} = .761$) nicht signifikant waren, erfüllten die Daten die Voraussetzungen für die Varianzanalyse. Die Varianzanalyse auf Unterschiede zwischen den drei Niveaus ergab eine Gesamtsignifikanz von $p < .05$ (Tab. 18). Die Effektstärke betrug $\eta^2_{Itemschwierigkeit} = .175$. Demnach erklärten die Niveaustufen 17.5 % der Varianz.

Tab. 18: Einfaktorielle Varianzanalyse mit der Itemschwierigkeit als abhängiger Variable und den Itemniveaus als unabhängiger Variable.

Quelle der Variation	Quadratsumme (SS)	Freiheitsgrade (df)	Varianz (MS)	F-Wert
between	4.462	2	2.231	3.936*
within	20.972	37	0.567	
total	25.434	39		

$p = .028$. * = signifikant ($p < .05$).

4.3.3 Beschreibung von Modellkompetenz im Querschnitt

Die Mittelwerte der PVs als Personenfähigkeitsschätzer stiegen mit der Jahrgangsstufe linear an. Der größte Sprung lag dabei mit .357 *logits* zwischen den Jahrgangsstufen 8 und 9 (Tab. 19; Abb. 23).

Tab. 19: Mittelwerte und Standardabweichungen der PVs für die Jahrgangsstufen 7 bis 10.

	Jgst. 7		Jgst. 8		Jgst. 9		Jgst. 10	
M_{PV1} (SD_{PV1})	.080	(.9386)	.312	(.8810)	.607	(.9093)	.638	(.8461)
M_{PV2} (SD_{PV2})	.040	(.9220)	.271	(.9195)	.616	(.9690)	.633	(.8883)
M_{PV3} (SD_{PV3})	.098	(.9645)	.262	(.8711)	.668	(.9786)	.705	(.9185)
M_{PV4} (SD_{PV4})	.083	(.9120)	.243	(.9170)	.642	(.8982)	.649	(.9335)
M_{PV5} (SD_{PV5})	.059	(.9967)	.300	(.9761)	.641	(.8990)	.688	(.9441)
M_{PV} (SD_{PV})	.072	(.5004)	.278	(.4720)	.635	(.4661)	.663	(.4530)

Die PV*s* sind logit-skaliert. Jgst. = Jahrgangsstufe.

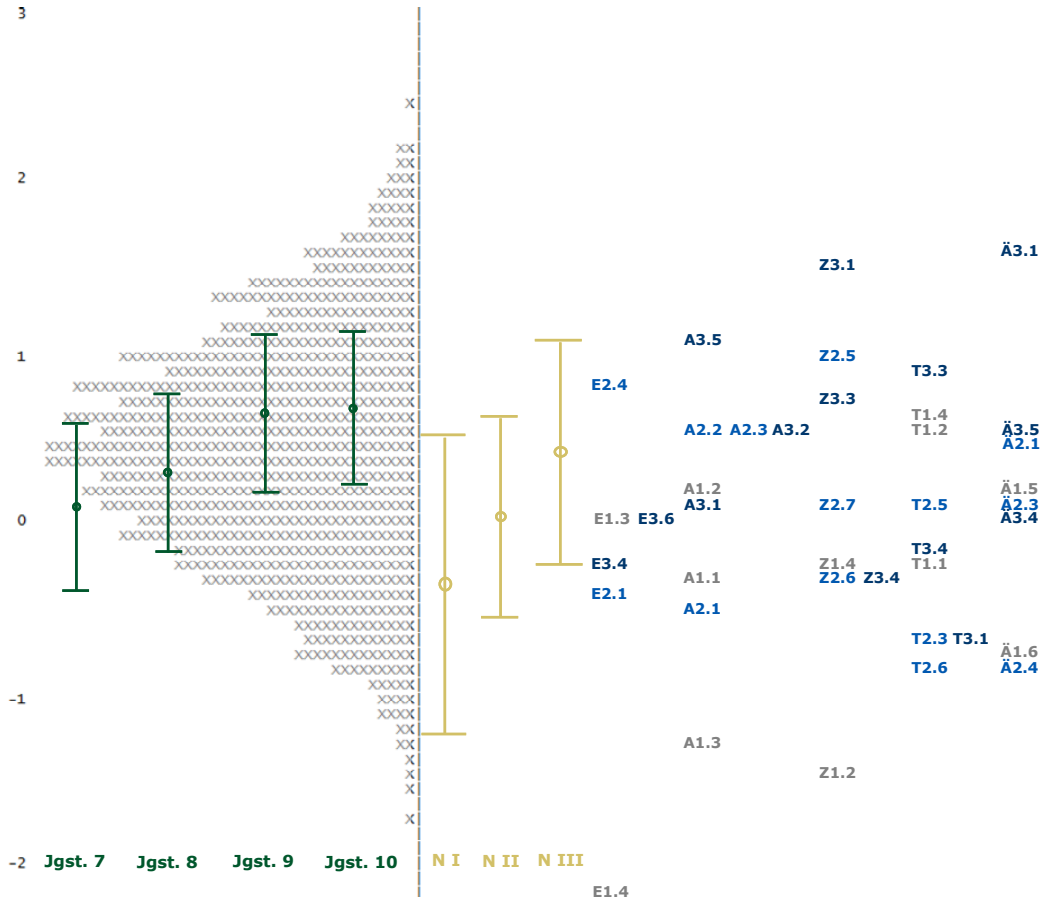


Abb. 23: Wright Map mit Personenfähigkeiten je Jahrgangsstufe (gemittelte PVs je Person) sowie einzelnen Items mit Itemschwierigkeiten[18] je Niveau als Boxplots. Der Kreis in den Boxplots zeigt jeweils den Mittelwert an, die Whisker (Linien) eine Standardabweichung um den Mittelwert. Jgst. = Jahrgangsstufe, N = Niveau.

In einer latenten Regression der Modellkompetenz auf die Jahrgangsstufe zeigte sich, dass diese Anteile der Varianz von Modellkompetenz erklärt (Tab. 20). Die Regressionskonstante B_0 stellt den Mittelwert der Personenfähigkeiten der Schülerinnen und Schüler in der Jahrgangsstufe 7 dar. Die Schülerin-

[18] Im Unterschied zu Abb. 11 liegt dieser Abbildung, auch für die Boxplots der Niveaus, eine itemzentrierte Skalierung zugrunde.

nen und Schüler verfügten in der Jahrgangsstufe 8 im Mittel über eine Modellkompetenz von .262 (= .067 + .195), in der Jahrgangsstufe 9 von .666 (= .067 + .599) und in der Jahrgangsstufe 10 von .673 (= .067 + .606). Unter Auspartialisierung der Jahrgangsstufe blieb eine Varianz von .433 im Vergleich zu .510[19] ohne diese Auspartialisierung. Die Jahrgangsstufe erklärte 15 % der Varianz in der Modellkompetenz der Schülerinnen und Schüler (R^2 = .15) und war mit einem F-Wert von 66.59 signifikant. Dies gilt ebenfalls für die Regressionskoeffizienten der einzelnen Jahrgangsstufen, die hochsignifikant waren (Tab. 20).

Tab. 20: Latente, unstandardisierte Regressionskonstante B_0, Regressionsgewichte B_I, Standardfehler SE_B sowie t-Werte der Jahrgangsstufen 7 bis 10 auf Modellkompetenz.

	B_0 bzw. B_I (SE_B)		β_I	t-Wert
Jgst. 7	.067	(.038)		
Jgst. 8	.195	(.054)	.296	3.61***
Jgst. 9	.599	(.055)	.910	16.89***
Jgst. 10	.606	(.055)	.921	11.01***

Jgst. = Jahrgangsstufe, ns = nicht signifikant, * = signifikant ($p < .05$), ** = sehr signifikant ($p < .01$), *** = hochsignifikant ($p < .001$).

Die zusätzliche Auspartialisierung des Alters führte zu einer Varianz von .416. Somit erklärte die Einbeziehung des Alters über den Effekt der Jahrgangsstufe hinaus weitere 3 % (R^2 = .18) der Varianz. Das Alter trug nicht signifikant zur Erklärung von Modellkompetenz bei.

[19] Dieser Wert stimmt nicht mit der in Kapitel 4.3.1 dargestellten Varianz der eindimensionalen Skalierung überein, weil dort die Metrik der Skala durch die Fixierung der Personenfähigkeiten festgelegt wurde. Da die latenten Regressionsanalysen auf Informationen über die Personen abzielen, wird hier die Metrik über die Itemschwierigkeiten definiert.

Tab. 21: Latente, unstandardisierte Regressionskonstante B_0, Regressionsgewichte B_I, Standardfehler SE_B sowie t-Werte der Jahrgangsstufen 7 bis 10 und des Alters auf Modellkompetenz.

	B_0 bzw. B_I (SE_B)	β_I	t-Wert
Jgst. 7	.564	(.388)	
Jgst. 8	.247	(.064)	.375	3.86***
Jgst. 9	.661	(.081)	1.00	8.16***
Jgst. 10	.740	(.112)	1.12	6.61***
Alter	-.039	(.030)	-.059	-1.30 ns

Jgst. = Jahrgangsstufe, ns = nicht signifikant, * = signifikant ($p < .05$), ** = sehr signifikant ($p < .01$), *** = hochsignifikant ($p < .001$).

4.3.4 Beziehungen von Modellkompetenz zu anderen Konstrukten

Tab. 22 zeigt die Mittelwerte der Korrelationen und Standardabweichungen je Variable. Bis auf die Korrelationen zwischen Modellkompetenz und Geschlecht bzw. Wissenschaftsverständnis sind alle Korrelationen auf einem Niveau von $p < .01$ signifikant. Die Korrelation mit dem Geschlecht ist nicht signifikant, die mit dem Wissenschaftsverständnis je nach PV und imputiertem Datensatz nicht signifkant bis zu signifikant auf einem Niveau von $p < .01$. Die vollständige Tabelle mit allen Werten ist in Anhang 10 zu finden.

Tab. 22: Mittelwerte und Standardabweichungen der Korrelationen der Personenfähigkeiten für Modellkompetenz mit weiteren Variablen.

	Geschlecht	allgemeine kognitive Fähigkeiten	Lesegeschwindigkeit	Leseverstehen	NOS	Chemieunterricht
$M_{Korrelation}$	.016	.351	.523	.486	-.072	.215
$SD_{Korrelation}$	.0039	.0004	.0059	.0001	.0115	.0003

Fortsetzung Tab. 22.

	Biologie-note	Chemie-note	Physiknote	Mathe-matiknote	Deutsch-note	Note erste Fremd-sprache
$M_{Korrelation}$	-.289	-.253	-.369	-.319	-.304	-.233
$SD_{Korrelation}$	.0003	.0004	.0004	.0002	.0004	.0004

4.4 Diskussion

Die Ergebnisse werden herangezogen, um Modellkompetenz empirisch für eine effiziente Förderung zu beschreiben. Inwiefern sich die theoriegeleitete Struktur von Modellkompetenz empirisch abbildet, wird anhand der Befunde zur Strukturierung (Kapitel 4.4.1) und Graduierung (Kapitel 4.4.2) diskutiert. Zu den Fragen, mit welchen Lernvoraussetzungen Lehrerinnen und Lehrer rechnen können und inwiefern Modellkompetenz erlernbar ist, gibt die querschnittliche Betrachtung von Performanzunterschieden zwischen den Jahrgangsstufen 7 bis 10 Aufschluss (Kapitel 4.4.3). Die Erlernbarkeit von Modellkompetenz wird außerdem neben der Domänenspezifität anhand der Beziehungen von Modellkompetenz zu anderen Konstrukten diskutiert (Kapitel 4.4.4). Diese Beziehungen werden zur konvergenten Validierung von Modellkompetenz herangezogen. Wenn die methodische Umsetzung die Interpretation einzelner Ergebnisse beeinflusst, ist dies in den Text integriert. Eine übergreifende Methodendiskussion fasst diese Punkte abschließend zusammen (Kapitel 4.4.5).

4.4.1 Überprüfung der Strukturierung von Modellkompetenz

Die empirische Strukturierung von Modellkompetenz ermöglicht Aussagen über deren Dimensionalität, deren Domänen- bzw. Kontextspezifität sowie Verbindungen zwischen den Teilkompetenzen.

Strukturierung in Dimensionen und Teilkompetenzen

Um Modellkompetenz empirisch zu strukturieren, wurden ein ein-, zwei- sowie fünfdimensionales Strukturmodell sowie zwei Dummy-Modelle miteinander verglichen. Zwischen der *Deviance* und somit der Passung der drei theoriegeleiteten Strukturmodelle gab es keine signifikanten Unterschiede. Sie beschreiben demnach gleichermaßen gut die Daten. Darüber hinaus unterschied sich die Passung der Dummy-Modelle kaum von den entsprechenden theoriegeleiteten Modellen. Auch die Informationskriterien zeigten keine großen Unterschiede zwischen den Strukturmodellen. Sie deuten mit dem kleinsten Wert für das eindimensionale Modell darauf hin, dass dieses Modell die Daten am besten beschreibt. Der χ^2-Test, der Vergleich der theoriegeleiteten mit den Dummy-Modellen sowie die Informationskriterien sprechen damit nicht gegen eine eindimensionale Struktur (**H0**) und liefern keine Indizien für eine höhere Dimensionalität (**H1** oder **H2**). Die Korrelationen zwischen den Dimensionen ‚Kenntnisse über Modelle' und ‚Modellbildung' sowie zwischen den Teilkompetenzen waren mit $.499 < r < .763$ jedoch für messfehlerbereinigte Korrelationen nicht sehr hoch (Carstensen et al., 2007). Dies wäre ein Indiz dafür, dass eine eindimensionale Lösung nicht optimal ist und mehrere Dimensionen angenommen werden sollten. Die Datenbasis für die Schätzung der Zusammenhänge war allerdings schwach, da die Varianz der Personenparameter niedrig war (Kapitel 3.1.6). Dies führt zu niedrigeren Korrelationen zwischen den Dimensionen, weil bei geringer Varianz die Korrelationen unterschätzt werden. Diese sind deshalb nicht belastbar zu interpretieren.

Die heterogenen Befunde zur Strukturierung von Modellkompetenz sind somit möglicherweise auf die Datenbasis mit geringer Varianz und Reliabilität sowie einer geringen Abdeckung der Kovarianzmatrix zurückzuführen. Mit dem gleichen Untersuchungsdesign sowie ähnlichen Varianzen und Reliabilitäten kamen jedoch Krell und Krüger (2011) mit FC-Items zu signifikanten Unterschieden zwischen den theoriegeleiteten Strukturmodellen. Diese sprechen für die Annahme des fünfdimensionalen Modells. Demnach ist das Design offenbar grundsätzlich geeignet, eindeutige Aussagen zu treffen. Der uneindeutige Befund mit den MC-Items könnte durch eine mangelnde Qualität der Items

bedingt sein, die die Aussagekraft der Ergebnisse stark beeinträchtigen kann. Dies ist für die hier vorgestellte Studie aber unwahrscheinlich, da gezeigt werden konnte, dass die MC-Items Modellkompetenz angemessen repräsentieren (Kapitel 3.1.4, 3.2) und als Indikator für dieses Konstrukt interpretierbar sind (Kapitel 3.3). Dass Krell und Krüger (2011) mit einer vergleichbaren Datenbasis im Modellvergleich eindeutige Befunde erzielten, stützt demnach die Aussagekraft der vorliegenden Datenbasis und spricht dafür, dass die Hinweise auf eine eindimensionale Struktur von Modellkompetenz aus den hier analysierten Daten durchaus relevant sind.

Domänen- bzw. Kontextspezifität

Die unterschiedlichen Ergebnisse dieser beiden Aufgabenformate sind möglicherweise darauf zurückzuführen, dass die FC-Items weniger stark in Situationen eingebunden sind und die Antwortmöglichkeiten dort auf einer abstrakteren Ebene formuliert sind. Mit ihnen wird stärker ein abstraktes Modellverständnis erhoben. In diesem Zusammenhang weisen Sins et al. (2009) darauf hin, dass Schülerinnen und Schüler epistemologische Vorstellungen eher angebunden an einen Kontext, aber evtl. nicht als abstrakte Standpunkte formulieren können oder abstrakte Formulierungen nach Hörensagen ankreuzen. Auch Leisner und Mikelskis (2004) sowie Leisner-Bodenthin (2006) differenzieren zwischen domänenübergreifender und domänenunabhängiger Modellkompetenz. Für das Wissenschaftsverständnis, das Upmeier zu Belzen und Krüger (2010) der Graduierung von Modellkompetenz in Anlehnung an Grosslight et al. (1991) zugrunde legen (Kapitel 2.3.3), zeigten Urhahne et al. (2011), dass kontextspezifische Vorstellungen mit kontextunabhängigen in Zusammenhang stehen und eine Vertrautheit mit vielen Kontexten mit einem elaborierten Wissenschaftsverständnis einhergeht. Eine Differenzierung unterschiedlich kontextspezifischer Vorstellungen erscheint nach den Befunden dieser Studien relevant. Entsprechend könnte es sein, dass die Bearbeitung der MC-Items, die verglichen mit den FC-Items domänen- und kontextspezifischer sind, anders strukturiert ist als die Bearbeitung der FC-Items.

Verbindungen zwischen Teilkompetenzen

Die Hinweise auf eine eindimensionale Struktur des konkreten Umgangs mit Modellen wären möglicherweise dadurch zu erklären, dass die Teilkompetenzen im konkreten Umgang eng miteinander verflochten sind. Dies würde empirisch die theoriebasierte Einschätzung von Schwarz et al. (2009) unterstützen, dass Schülerinnen und Schüler beim konkreten Umgang mit Modellen immer auf Vorstellungen aus verschiedenen inhaltlichen Bereichen zurückgreifen. Auf Zusammenhänge zwischen den Teilkompetenzen weisen die latenten Korrelationen zwischen ihnen hin. Da sie allerdings möglicherweise nicht in gleichem Ausmaß unterschätzt werden, könnte sich das Bild bei einer höheren Varianz verändern. Die Deutung im Folgenden ist deshalb als hypothetisch und vorläufig zu verstehen.

Eigenschaften von Modellen, Alternative Modelle, Zweck von Modellen

Auch Verbindungen zwischen den Kategorien zu Vorstellungen zu Modellen legen eine enge Verknüpfung zwischen Bereichen von Modellkompetenz nahe. Dies betrifft z. B. ‚Alternative Modelle' und ‚Eigenschaften von Modellen' (Grosslight et al., 1991): Der Gedanke, dass alternative Modelle aufgrund verschiedener Materialien (Niveau I) oder Blickwinkel (Niveau II) existieren, verweist auf ein einfaches Modellverständnis im Bereich ‚Eigenschaften von Modellen' (Grosslight et al., 1991). Wenn Schülerinnen und Schüler alternative Modelle mit verschiedenen Aspekten oder Detailliertheitsgraden begründen (Niveau II), verweist dies auf ein Verständnis von Modellen als vereinfachte Repräsentationen (Niveau II). Verschiedene Hypothesen als Begründung für alternative Modelle (Niveau III) deuten auf eine Auffassung von Modellen als theoretische Rekonstruktionen hin (Niveau III). Einen Zusammenhang zwischen den Teilkompetenzen ‚Eigenschaften von Modellen' und ‚Alternative Modelle' legt auch die latente Korrelation zwischen ihnen nahe (r = .679). Alternative Modelle können außerdem durch verschiedene Zwecke dieser Modelle zur Kommunikation begründet werden (Grosslight et al., 1991). Zusammenhänge zwischen ‚Eigenschaften von Modellen', ‚Alternativen Modellen' und dem ‚Zweck von Modellen' wurden bereits bei den Experten-Ratings (Kapitel 3.1.4,

3.2) und den kognitiven Prozessen der Schülerinnen und Schüler bei der Itembearbeitung deutlich (Kapitel 3.3). Der ‚Zweck von Modellen' hängt mit diesen beiden Teilkompetenzen in den hier diskutierten Daten allerdings in geringerer Höhe zusammen als mit den Teilkompetenzen im Bereich ‚Modellbildung'. Die vorliegenden Ergebnisse stützen demnach nicht seine Einordnung als deklaratives Wissen und somit in die Kenntnisse über Modelle (Leisner-Bodenthin, 2006; **H15b**). Inwiefern Niveau I dieser Teilkompetenz mit den ‚Eigenschaften von Modellen' und ‚Alternativen Modellen' zusammenhängt (**H15a**), kann anhand dieser Befunde nicht beurteilt werden. Hierfür müssten Strukturgleichungsmodelle spezifiziert werden, für deren Berechnung die Abdeckung der Kovarianzmatrix im hier verwendeten Testheftdesign (Kapitel 4.1) nicht ausreicht.

An den latenten Korrelationen fällt außerdem auf, dass die Teilkompetenz ‚Eigenschaften von Modellen' vergleichsweise eng mit allen anderen Teilkompetenzen zusammenhing ($.591 < r < .682$). Dies würde nahelegen, sie als globalen Faktor anzunehmen, auf dem alle anderen Teilkompetenzen aufbauen. Zu Items in offenem Antwortformat formulierten Schülerinnen und Schüler konsistente Perspektiven auf Modelle. Dies deuten Grünkorn et al. (in Vorb.) als Hinweis auf stabile kognitive Strukturen in diesem Bereich, die sie in anderen Bereichen nicht finden. Auch das könnte als ein Indiz für eine grundlegende Rolle der ‚Eigenschaften von Modellen' interpretiert werden. Inhalt dieser Teilkompetenz ist auf allen drei Niveaus eine Perspektive auf Modelle *von* etwas (Mahr, 2008a; Upmeier zu Belzen & Krüger, 2010) und somit die Beziehung zwischen Modell und Original. Diese wurde sowohl in den Experten-Ratings (Kapitel 3.1.4, 3.2) als auch beim lauten Denken (Kapitel 3.3) als Vergleich von Modell und Original prominent thematisiert. Die gesamte Teilkompetenz ‚Eigenschaften von Modellen' hat im Vergleich zum ‚Testen von Modellen', Niveau II, einen stärker deklarativen Fokus. Hier steht weniger die Frage nach der Einsetzbarkeit und Anwendbarkeit des Modells für einen bestimmten Zweck im Vordergrund, sondern die grundsätzliche Beziehung, die Modell und Original haben können. Dies wäre eine Erklärung, warum insbesondere diese Teilkompetenz eine zentrale Rolle spielen könnte.

Zweck, Testen und Ändern von Modellen

Die latenten Korrelationen zwischen den Teilkompetenzen ‚Zweck' und ‚Testen von Modellen' (r = .616) sowie zwischen ‚Testen' und ‚Ändern von Modellen' (r = .763) waren höher als die zwischen anderen Teilkompetenzen. Beim ‚Zweck', ‚Testen' und ‚Ändern von Modellen' bildeten sich damit möglichweise empirisch enge Zusammenhänge ab, die theoriegeleitet erklärt werden könnten: Upmeier zu Belzen und Krüger (2010) orientieren sich bei der Zuordnung dieser drei Teilkompetenzen zu einer Dimension an der Konzeption von Justi und Gilbert (2002). Dort ist der Zweck im *model of modelling* der Startpunkt der Anwendung eines Modells. Ausgehend davon werden gedankliche und empirische Tests durchgeführt, die zu einer Änderung des Modells führen können. Die latenten Korrelationen zwischen diesen Teilkompetenzen könnten darauf hindeuten, dass sich diese zeitliche Abfolge in der Struktur von Modellkompetenz ausdrückt: Der ‚Zweck von Modellen' hängt unmittelbar mit dem ‚Testen von Modellen' zusammen, das wiederum einen unmittelbaren Zusammenhang mit dem ‚Ändern von Modellen' aufweist.

Auch die Kategorisierung von Vorstellungen zu Modellen legt nahe, dass der Zweck mit dem Testen von Modellen zusammenhängt: Schülerinnen und Schüler sowie Lehrerinnen und Lehrer nennen als Zwecke, die mit Modellen verbunden sein können, u. a. dass ein Modell ein Original bei Tests ersetzen kann (Crawford & Cullin, 2005) und mit Modellen Vorhersagen getestet werden können (Crawford & Cullin, 2005; Justi & Gilbert, 2003). Grünkorn et al. (in Vorb.) stellten ebenfalls fest, dass die zweckgebundene Entwicklung von Modellen sich insbesondere in Kategorien der Teilkompetenzen ‚Testen' und ‚Ändern von Modellen' niederschlägt.

Der Zusammenhang zwischen dem ‚Testen' und ‚Ändern von Modellen' in den latenten Korrelationen korrespondiert mit den Ergebnissen des lauten Denkens (Kapitel 3.3.3), dass Niveau II und III dieser Teilkompetenzen möglicherweise miteinander zusammenhängen (**H13**, **H14**). Enge Zusammenhänge zwischen Antworten aus Items zum ‚Testen' und ‚Ändern von Modellen' fanden auch Grünkorn et al. (in Vorb.), die beschreiben, dass Schülerinen und Schü-

ler in diesen Teilkompetenzen häufig ähnliche Perspektiven auf Modelle und die Modellbildung formulieren.

Grünkorn et al. (in Vorb.) stellen neben Kategorien, die dem Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) zugeordnet wurden, zwei Kategorien dar, die in Items zu allen Teilkompetenzen genannt wurden: ‚Modelle als Mittel zur Verständlichkeit und Kommunizierbarkeit' sowie ‚Modelle als Mittel zur Zugänglichkeit'. Beide Kategorien traten vorwiegend in Verbindung mit einem medialen Verständnis von Modellen auf. Auch die Existenz solcher quer zum Kompetenzmodell liegenden Kategorien könnte auf enge Zusammenhänge zwischen Teilkompetenzen und Niveaus hindeuten.

Diese vielfältigen Bezüge zwischen Teilkompetenzen sprechen dafür, dass Schwarz et al. (2009) Vorstellungen im konkreten Umgang mit Modellen zurecht für sehr vernetzt halten. Ein globaler Bezug, wie Schwarz et al. (2009) ihn herstellen, ist jedoch in der Unterrichtsgestaltung schwer aufzugreifen. Hier bietet das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) mehr Anknüpfungspunkte. Aus diesem Grund wäre weiterhin eine explizite Modellierung der Bezüge interessant, um über einzelne Bezüge zwischen Kategorien von Vorstellungen zu Modellen hinaus spezifische Zusammenhänge zwischen Teilkompetenzen aufzuklären, auch wenn diese möglicherweise so eng sind, dass sich empirisch keine Dimensionen differenzieren lassen (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003).

Weiterführende Studien mit einem Design, das die Testung detaillierter Hypothesen im Rahmen von Strukturgleichungsmodellen erlaubt, könnten somit die Indizien für eine eindimensionale Struktur des konkreten Umgangs mit Modellen sowie enge Verbindungen zwischen den Teilkompetenzen weiter verfolgen. Um hierfür eine Abdeckung der Kovarianzmatrix zu erreichen, die solche Analysen erlaubt, müssten weniger Bereiche von Modellkompetenz gleichzeitig untersucht oder mehr Items je Person vorgegeben werden. Dabei sollte darauf geachtet werden, dass die Personen nicht so viele Items bekommen, dass sie sie nicht mehr konstruktgemäß bearbeiten, da sich bereits in dieser

Studie ein *Optimizing-Satisficing*-Problem zeigte (Jonkisz, Moosbrugger & Brandt, 2012; Kapitel 3.1.6).

4.4.2 Überprüfung der Graduierung von Modellkompetenz

Da die MC-Items jeweils ein Niveau einer Teilkompetenz abbilden, ermöglichen die Daten Aussagen über die globale Graduierung in Niveaus. Über die Graduierung einzelner Teilkompetenzen können aufgrund der starken Streuung der Itemschwierigkeiten über die Niveaus nur eingeschränkt Aussagen getroffen werden.

Graduierung in Niveaus

Zur empirischen Überprüfung der Graduierung von Modellkompetenz wurden die Itemschwierigkeiten herangezogen. Diese streuten häufig je Teilkompetenz und Niveau relativ breit (vgl. Abb. 11, Abb. 22). Die Hinweise, die daraus abgeleitet werden, sind deshalb eher als hypothesengenerierend zu verstehen.

Mit Abweichungen in einzelnen Bereichen unterschieden sich die mittleren Itemschwierigkeiten aus der IRT-Skalierung signifikant mit den Niveaustufen (Tab. 18). Ein $\eta^2_{Itemschwierigkeit}$ von .175 stellt nach Sedlmeier und Renkewitz (2008) einen großen Effekt dar. Die Niveaus erklären demnach einen substanziellen Anteil der Varianz der Itemschwierigkeit. Deshalb kann angenommen werden, dass die Niveaus ansteigende Anforderungen an die Schülerinnen und Schüler abbilden (**H3**). Entsprechend erscheint es sinnvoll, drei Niveaus zu differenzieren und anzunehmen, dass die methodische Nutzung von Modellen *für* etwas (Mahr, 2008a) höhere Anforderungen an Schülerinnen und Schüler stellt als die mediale von Modellen *von* etwas (Mahr, 2008a). Dies korrespondiert mit der Konzeption der Graduierung, die Schwarz et al. (2009) vornehmen, und Ergebnissen der Studie von Grosslight et al. (1991), an der sich das Kompetenzmodell von Upmeier zu Belzen und Krüger (2010) in der Graduierung von Modellkompetenz orientiert. Diese Parallelen sprechen dafür, dass die Niveaus im Kompetenzmodell insgesamt sinnvoll definiert wurden. Der Befund, dass eine mediale und eine methodische Perspektive voneinander

abgrenzbar sind, stützt den Ansatz von Henze et al. (2007), sie mit Blick auf verschiedene Ziele naturwissenschaftlichen Unterrichts (Hodson, 1993) gezielt einzusetzen.

Da die Abstände zwischen den Niveaus empirisch nicht sehr groß sind, erscheint eine weitere Ausdifferenzierung in vier Niveaus, wie Crawford und Cullin (2005) sie vornehmen, nicht sinnvoll. Auch wenn dieses Ergebnis nicht besagen muss, dass Schülerinnen und Schüler über kohärente ontologische und epistemologische Sichtweisen auf Modelle verfügen, geben die hier erzielten Befunde keinen Anlass dazu, die Annahme von unterschiedlich komplexen Niveaus abzulehnen.

Als Merkmale, die neben der Zuordnung zu den Niveaus schwierigkeitserzeugend waren, kommen z. B. Textlänge (Prenzel et al., 2002), Wortschatz (Hartig & Klieme, 2006; Rost, 2004) und fachlicher Inhalt sowie Situationen in den Items (Prenzel et al., 2002; Schecker & Parchmann, 2006) infrage (vgl. Kapitel 3.1.2). Auch wenn die Antwortmöglichkeiten auf der Grundlage von Schülerantworten formuliert wurden, die Position des Attraktors variiert wurde, die Items von Expertinnen und Experten diskutiert wurden und Formulierungen wie „immer", „nie" usw. vermieden wurden, ist nicht ganz auszuschließen, dass das logische Schließen die Itembearbeitung beeinflusst. Es erscheint deshalb lohnenswert, in anschließenden Studien zu prüfen, welche dieser Variablen weitere Anteile der Varianz in den Itemschwierigkeiten erklären und wie groß ihr Einfluss jeweils ist.

Graduierung einzelner Teilkompetenzen

Da die Itemschwierigkeiten stark über die Niveaus streuen, ist auf Grundlage der hier diskutierten Daten die Graduierung einzelner Teilkompetenzen nicht sicher beurteilbar. Es fällt jedoch auf, dass in der Teilkompetenz ‚Testen von Modellen' die meisten einfachen Items Niveau II zuzuordnen sind und die Items auf Niveau III im Vergleich zum ‚Zweck' und ‚Ändern von Modellen' einfacher sind. Vor dem Einsatz der Items zur Überprüfung der Graduierung wurde geprüft, inwiefern sie das zugrunde liegende Kompetenzmodell repräsen-

tieren (vgl. Kapitel 3.1.4, 3.2) und inwiefern die kognitiven Prozesse, die während ihrer Bearbeitung ablaufen, sich auf die intendierten Teilkompetenzen beziehen (vgl. Kapitel 3.3). Daher ist es unwahrscheinlich, dass diese abweichende Reihung der Itemschwierigkeiten auf die Items selbst zurückzuführen ist. Hinzu kommt, dass Krell (2012) mit FC-Items und Grünkorn et al. (in Vorb.) mit Items in offenem Antwortformat zu einem vergleichbaren Befund kamen. Da diese Items Modellkompetenz durch die Bewertung abstrakter Aussagen über Modelle bzw. in offenem Antwortformat erfassen und somit einen anderen methodischen Ansatz als das hier vorgestellte Projekt verfolgen, ist anzunehmen, dass sich dieser Befund auf die theoretische Grundlage zurückführen lässt.

Die Niveau II-Items der Teilkompetenz ‚Testen von Modellen' beziehen sich darauf, über die Beurteilung der Passung zwischen Modell und Original ein Modell für eine bestimmte Fragestellung auszuwählen (Kapitel 3.1.3, Anhang 2). Dies ähnelt stark dem Vergleich von Modell und Original, der sich bereits in den Experten-Ratings (Kapitel 3.1.4, 3.2) und beim lauten Denken (Kapitel 3.3) als zentrales Konzept herausstellte. Dieser Vergleich ist unter medialer Perspektive mit Blick auf die Fragen relevant, inwiefern Modelle Originale repräsentieren (Oh & Oh, 2011), sowie unter methodischer Perspektive, welche Merkmale das Modell für die vorgegebene Fragestellung umfassen sollte (Intentionalität von Modellen, Giere 2009; Stachowiak, 1983) und inwiefern hypothetisch Rückschlüsse vom Modell auf das Orginal gezogen werden können (Modell für etwas, Mahr, 2008a). Die Passung zwischen Modell und Original beinhaltet diesen Vergleich und geht in dessen Beurteilung über ihn hinaus. Möglicherweise ist dieses Konzept damit ebenfalls so grundlegend, dass Schülerinnen und Schüler darin so geübt sind, dass ihnen Aufgaben in diesem Bereich sehr leicht fallen. Dies stimmt mit dem Parallelisieren als prominenter Vorstellung bei Schülerinnen und Schülern überein (Grünkorn et al., in Vorb.; Trier & Upmeier zu Belzen, 2009).

Eine denkbare Erklärung für die geringe Schwierigkeit der Niveau III-Items beim ‚Testen von Modellen' im Vergleich zu denen beim ‚Zweck' sowie ‚Ändern von Modellen' wäre, dass Schülerinnen und Schüler das experimentelle Testen

von Hypothesen mit Originalen so gut kennen, dass ihnen die Bearbeitung entsprechender Aufgaben mit Modellen (Niveau III, Upmeier zu Belzen & Krüger, 2010) leichter fällt als die Reflexion darüber, welche Vorhersagen mit einem Modell getroffen werden (‚Zweck von Modellen', Niveau III) oder wie eine falsifizierte Hypothese sich auf ein Modell auswirkt (‚Ändern von Modellen', Niveau III). Dieses Ergebnis steht im Kontrast zu prominenten Vorstellungen zu Modellen, bei denen das Testen von Ideen und die Formulierung und Prüfung von Vorhersagen im Hintergrund stehen (Grosslight et al., 1991; Treagust et al., 2002). Dieser Unterschied könnte dadurch bedingt sein, dass sowohl Grosslight et al. (1991) als auch Treagust et al. (2002) Vorstellungen zu Modellen erhoben haben, ohne Kontexte zu verwenden. In der Bearbeitung der domänen- und kontextspezifischen MC-Items der vorliegenden Arbeit drückte sich im Vergleich eine elaboriertere Perspektive aus. Möglicherweise deutet dies darauf hin, dass Sins et al. (2009) richtig vermuten, dass Schülerinnen und Schüler epistemologische Vorstellungen in einem Kontext, aber evtl. nicht als abstrakten Standpunkt formulieren können. Auch Urhahne et al. (2011) stellten fest, dass sich kontextspezifische und -unabhängige Vorstellungen im Wissenschaftsverständnis überlappen, aber voneinander unterschieden werden können, und kontextspezifisches Wissen nicht zu generalisierten Vorstellungen führen muss.

Für den ‚Zweck' und das ‚Ändern von Modellen' stimmen die Ergebnisse jedoch mit kontextunabhängigen Formaten überein: Dort werden die Nutzung von Modellen zur Entwicklung von Theorien bzw. die Änderung eines Modells als Teil des Forschungsprozesses ebenfalls kaum genannt (AAAS, o. J.; Grosslight et al., 1991; Schwarz & White, 2005; Treagust et al., 2002; Trier & Upmeier zu Belzen, 2009). Schülerantworten auf kontextspezifische Items in offenem Antwortformat bezogen sich ebenso nur selten auf die Nutzung von Modellen zum Testen konkreter Hypothesen oder falsifizierte Hypothesen als Grund für die Änderung eines Modells (Grünkorn et al., in Vorb.). Möglicherweise bestehen damit nicht in allen Bereichen Unterschiede zwischen kontextspezifischen und –unabhängigen Formaten. Eine gezielte Erweiterung des Itempools könnte sowohl über den Einfluss von Kontextmerkmalen als auch über die Graduierung der einzelnen Teilkompetenzen Aufschluss geben.

4.4.3 Beschreibung von Modellkompetenz im Querschnitt

Im Vergleich der Jahrgangsstufen 7 bis 10 zeigten sich sowohl anhand der Unterschiede zwischen den Jahrgangsstufen Hinweise auf die Erlernbarkeit von Modellkompetenz als auch auf die Ausprägung von Modellkompetenz, über die Schülerinnen und Schüler in diesen Jahrgangsstufen verfügen.

Erlernbarkeit von Modellkompetenz

Die Betrachtung der Modellkompetenz-Personenfähigkeiten von Jahrgangsstufe 7 bis 10 zeigte, dass diese in Abhängigkeit von der Jahrgangsstufe signifikant höher wurden und mit etwa 15 % substanzielle Anteile der Varianz von Modellkompetenz erklären. Da somit in höheren Jahrgangsstufen tatsächlich höhere Leistungen erbracht wurden, ist **H4** anzunehmen. Das Alter erklärt darüber hinaus weitere 3 % und trägt nicht signifikant zur Erklärung der Personenfähigkeiten bei. Dementsprechend hat das Alter keinen substanziellen Effekt auf die Personenfähigkeiten, so dass auch **H5** angenommen werden kann. Die Befunde sprechen somit nicht dagegen, Modellkompetenz als erlern- und vermittelbar anzunehmen. Sie ist demnach zutreffend als Kompetenzkonstrukt definiert (vgl. Hartig & Klieme, 2006; Klieme & Hartig, 2007; Klieme & Leutner, 2006; Klieme, Maag-Merki et al., 2007; Koeppen et al., 2008; McClelland, 1973).

Sowohl die Niveaus (Kapitel 4.3.2) als auch Unterschiede in den Personenfähigkeiten (Kapitel 4.3.3) bilden sich statistisch ab. Auffällig ist jedoch, dass die Niveaus tendenziell unterhalb der Personenfähigkeiten liegen, so dass in Frage steht, inwiefern sie zur Beschreibung von Personenunterschieden geeignet sind. Da aber neben der a priori Zuordnung der Items zu den Niveaus auch andere schwierigkeitserzeugende Merkmale zu einem solchen Effekt führen könnten, kann hierzu keine abschließende Aussage getroffen werden.

Eine detailliertere Betrachtung der Ergebnisse zu Unterschieden zwischen den Jahrgangsstufen zeigt, dass der Abstand zwischen ihnen insgesamt mit .591 logits nicht sehr groß war. Dies ist vermutlich durch die insgesamt geringe Varianz (vgl. Kapitel 3.1.6) zu erklären, durch die keine sehr großen Unter-

schiede zu erwarten sind. Der Sprung von der Jahrgangsstufe 8 zu 9 und die geringen Unterschiede zwischen den Jahrgangsstufen 7 und 8 bzw. 9 und 10 legen nahe, dass der Unterricht sich nicht innerhalb, aber zwischen diesen Doppeljahrgangsstufen unterscheidet. Um zu klären, inwiefern Lehrerinnen und Lehrer in den beiden Doppeljahrgangsstufen Modelle unterschiedlich einsetzen, wäre es lohnenswert, die Lerngelegenheiten der Schülerinnen und Schüler zu untersuchen. Dies könnte außerdem darüber Aufschluss geben, welche Lernangebote insbesondere zur Kompetenzentwicklung beitragen (vgl. Patzke & Upmeier zu Belzen, 2012). Außerdem könnte die Erhebung der Lerngelegenheiten die Frage klären, inwiefern die geringe Varianz in der Modellkompetenz der Schülerinnen und Schüler dadurch bedingt war, dass diese nur wenige Lerngelegenheiten haben, um Modellkompetenz zu entwickeln. Es ist darüber hinaus zu beachten, dass die Items für diese Stichprobe eher einfach waren (vgl. Kapitel 3.1.6). Auch wenn der Deckeneffekt nicht so groß ist, dass die Trennschärfe der Items stark beeinträchtigt wäre (vgl. Kapitel 3.1.6) oder mit den Items keine Personenunterschiede beschrieben werden könnten (vgl. Kapitel 4.3.3), könnte der eher leichte Itempool die Ergebnisse beeinflussen. Eine Datenerhebung mit einem Itempool, der um zusätzliche schwierige Items erweitert wurde, könnte deshalb möglicherweise zu differenzierteren Befunden kommen.

Verfügbare Modellkompetenz der Schülerinnen und Schüler

Ein Vergleich der Personenfähigkeiten über die Jahrgangsstufen mit den Itemschwierigkeiten zeigt außerdem, dass Schülerinnen und Schüler in der Jahrgangsstufe 7 vor allem über Modellkompetenzen im medialen Bereich (Niveau I bis II, Upmeier zu Belzen & Krüger, 2010) verfügten. Bis zur Jahrgangsstufe 10 erwerben Schülerinnen und Schüler offenbar zunehmend Kompetenzen im methodischen Bereich – zumindest für die hier befragten Schülerinnen und Schüler war ein Unterschied in der Modellkompetenz festzustellen. Die mediale Perspektive auf Modelle, die vor allem bei den jüngeren Schülerinnen überwog, zählt nach Driver et al. (1996) sowie Stephens et al. (1999) nicht zum *model-based reasoning*. Erklärungen beziehen sich dort auf die Schlussfolgerung vom Modell als Original und somit nach Mahr (2008a)

ausschließlich auf das Modell *für* etwas (vgl. Kapitel 2.2.3). Das Ergebnis von Stephens et al. (1999), dass dies bei Schülerinnen und Schüler selten vorkommt, stimmt demnach mit den Befunden der vorliegenden Untersuchung überein.

Auffällig ist, dass Niveau III-Items beim ‚*Testen von Modellen*' einfacher zu sein scheinen als beim ‚*Zweck*' oder ‚*Ändern von Modellen*'. Diese drei Teilkompetenzen beinhalten auf Niveau III eine methodische Perspektive auf Modelle (Upmeier zu Belzen & Krüger, 2010). In Bezug auf die Graduierung von Modellkompetenz wurde bereits diskutiert (Kapitel 4.4.2), dass Schülerinnen und Schülers das ‚Testen von Modellen' möglicherweise deshalb vergleichsweise leicht fällt, weil sie diese Vorgehensweise vom Experimentieren mit Originalen kennen. Das Ergebnis, dass Schülerinnen und Schüler in diesem Bereich offenbar auf kontextabhängige Items elaborierter antworten, lässt sich nicht auf den ‚Zweck' und das ‚Ändern von Modellen' übertragen. Neben der Frage, inwiefern Schülerinnen und Schüler epistemologische Vorstellungen in einem Kontext, aber evtl. nicht als abstrakten Standpunkt formulieren können (Sins et al., 2009) stellt sich hier deshalb die Frage, welche Lerngelegenheiten sie in diesen Teilkompetenzen haben. Dass Schülerinnen und Schülern beim ‚Testen von Modellen' Items, die eine methodische Perspektive verlangen, häufiger lösen, könnte mit einer größeren Vertrautheit hiermit erklärt werden. Es ist davon auszugehen, dass Schülerinnen und Schüler das experimentelle Testen von Hypothesen mit Originalen kennen, während entsprechende Änderungen von Modellen oder Hypothesen als Grundlage von Modellen weniger häufig im Schulalltag auftreten dürften. Auch hier würde die Erfassung der Lerngelegenheiten bei einer Interpretation helfen.

Inwiefern die beschriebenen Unterschiede zwischen den Jahrgangsstufen tatsächlich als eine Kompetenzentwicklung im Sinne von intraindividuellen Unterschieden zu verstehen und nicht auf zufällige interindividuelle Unterschiede zwischen den Schülerinnen und Schülern im Querschnitt zurückzuführen sind, muss in einer längsschnittlichen Untersuchung geklärt werden (vgl. Patzke & Upmeier zu Belzen, 2011, 2012).

4.4.4 Beziehungen von Modellkompetenz zu anderen Konstrukten

Die Beziehungen von Modellkompetenz zu anderen Konstrukten sind sowohl mit Blick auf ihre Erlernbarkeit und Domänenspezifität als auch auf die diskriminante und konvergente Validität von Modellkompetenz interpretierbar.

Erlernbarkeit und diskriminante Validität

Die Korrelation von Modellkompetenz mit allgemeinen kognitiven Fähigkeiten (M_r = .351, SD_r = .0004) spricht dafür, dass Modellkompetenz erlernbar ist und stützt damit die Hinweise, die auf Unterschieden zwischen den Jahrgangsstufen beruhen. Im Vergleich zu Studien von PISA (r = .68; Leutner et al., 2004) sind diese für Modellkompetenz geringer. Möglicherweise kommt dies dadurch zustande, dass die Korrelationen durch die geringe Varianz von Modellkompetenz in den vorliegenden Daten unterschätzt werden. Da hier wie in PISA messfehlerbereinigte Korrelationen berechnet wurden, kommt die unterschiedliche Höhe nicht durch den Einfluss von Messfehlern zustande (Carstensen et al., 2007). Die Abgrenzbarkeit von Modellkompetenz (**H6**) wird durch die hier diskutierten Daten nicht in Frage gestellt, was als Hinweis auf die diskriminante Validität dieses Konstrukts gewertet werden kann.

Die Modellkompetenz-Personenfähigkeiten, die auf Grundlage der Bearbeitung der MC-Items geschätzt wurden, korrelierte in erwartbarer Höhe mit Lesegeschwindigkeit (M_r = .523, SD_r = .0059) und Leseverständnis (M_r = .486, SD_r = .0001). Damit kann die Bearbeitung des hier entwickelten Modellkompetenz-Tests ähnlich wie die Bearbeitung anderer Tests (z. B. PISA; Leutner et al., 2004) von Lesefähigkeiten abgegrenzt werden. Dieser Befund stützt **H10**. Er spricht außerdem wie die Beziehung zu allgemeinen kognitiven Fähigkeiten dafür, dass Modellkompetenz sinnvoll von anderen Konstrukten abgegrenzt werden kann und somit dieses Konstrukt diskriminant valide ist. Dies erhöht die Wahrscheinlichkeit, dass die geringe Varianz in der vorliegenden Untersuchung auf die mangelnde Implementierung von Modellkompetenz in den Unterricht zurückzuführen ist.

Domänenspezifität und konvergente Validität

Die Zusammenhänge von Modellkompetenz mit Leistungen in naturwissenschaftlichen Fächern waren wie angenommen (**H7**) schwach bis mittel (-.253 < M_r < -.369). Dieses Ergebnis korrespondiert mit Ergebnissen zu Kompetenzen im Bereich naturwissenschaftlicher Untersuchungen (Wellnitz, 2012) sowie naturwissenschaftlicher Kompetenz (Schütte et al., 2007). Dies gilt jedoch entgegen **H7** ebenfalls für die Leistungen in sprachlichen Fächern ($M_{r\ Deutsch}$ = -.304; $M_{r\ Erste\ Fremdsprache}$ = -.233), so dass der Unterschied zwischen diesen Bereichen sich nicht wie angenommen abbildet. **H7** wird demnach verworfen.

Grube (2011) sowie Urhahne et al. (2008) berichten geringere Korrelationen zwischen wissenschaftlichem Denken bzw. Wissenschaftsverständnis und Schulnoten in naturwissenschaftlichen Fächern. Hierzu stellten zum einen Dübbelde, Mayer, Möller und von Aufschnaiter (2011) fest, dass angehende Biologielehrerinnen und -lehrer Schwierigkeiten bei der Diagnose von wissenschaftsmethodischen Kompetenzen haben. Zum anderen stellt sich die Frage, inwiefern Noten über das Fachwissen hinaus den Kompetenzbereich Erkenntnisgewinnung abbilden. Wenn z. B. im naturwissenschaftlichen Unterricht die Benotung vor allem mit Blick auf das Fachwissen erfolgt, ist eine höhere Korrelation der Schulnoten in naturwissenschaftlichen als in sprachlichen Fächern nicht unbedingt plausibel. Darüber hinaus ist es möglich, dass Bezugsgruppeneffekte die Schulnoten stark beeinflussen und dadurch abweichende Korrelationen zustande kommen. Trautwein et al. (2008) berichten u. a. substanzielle Unterschiede zwischen einzelnen Schulen in der Benotung von Schulleistungen, so dass ein Einfluss dieser Variable denkbar ist. Auch nichtkognitive Aspekte wie Fleiß und Anstrengungsbereitschaft könnten unterschiedlich stark in die Benotung einfließen (Trautwein et al., 2008).

Der Befund, dass die Schülerinnen und Schüler, die noch keinen Chemieunterricht hatten, etwas ausgeprägtere Personenfähigkeiten mit Blick auf Modellkompetenz hatten (M_r = .215, SD_r = .0003), ist überraschend und auf Grundlage der vorliegenden Daten nicht zu erklären. Da dieser Unterschied auf vier

Klassen von drei verschiedenen Schulen (9.33 % der Schülerinnen und Schüler) beruht, für die angenommen werden kann, dass sie noch keinen Chemieunterricht hatten, kann es sein, dass zufällig diese Schülerinnen und Schüler besonders leistungsstark waren. Inwiefern der Umgang mit Modellen eine Arbeitsweise ist, die über den Biologieunterricht hinaus generalisiert und auf den Chemieunterricht bezogen werden kann, wird in der Studie ‚Modellierung und Diagnose horizontaler Vernetzung im Chemie- und Biologieunterricht der Sekundarstufe I' (VerE; Nowak, Nehring, Upmeier zu Belzen & Tiemann, 2012) untersucht.

Anders als erwartet (**H9**) hängt Modellkompetenz mit dem Wissenschaftsverständnis nicht zusammen. Modellkompetenz ist zwar wie angenommen vom Wissenschaftsverständnis abzugrenzen, aber müsste erwartungsgemäß stärker sowie positiv mit ihm zusammenhängen, da die Graduierung auf dieses Konstrukt aufbaut. **H9** wird damit verworfen. Möglicherweise ist dies dadurch zu erklären, dass Modellkompetenz domänen- und kontextspezifisch erhoben wurde, während der verwendete Test zum Wissenschaftsverständnis kontextunabhängig ist.

Die Befunde zur konvergenten Validität zeichnen bislang kein klares Bild. Hier wäre vermutlich die Bearbeitung der Frage, inwiefern kontextspezifische und -unabhängige Items unterschiedliche Facetten von Konstrukten erheben, hilfreich.

Geschlecht

In Anlehnung an Studien zu naturwissenschaftlichen Kompetenzen (Grube, 2011; Klos et al., 2008; Koslowski, 1996; Prenzel, Schöps et al., 2007; Wellnitz, 2012) ist für Modellkompetenz maximal ein geringer Geschlechtereffekt zugunsten der Mädchen zu vermuten. Für Modellkompetenz wurde im vorliegenden Projekt kein Unterschied festgestellt, so dass **H8** durch die Daten gestützt wird. Darüber hinaus bieten die Daten keinen Anlass für Lehrerinnen und Lehrer, in der Förderung von Modellkompetenz zwischen Schülerinnen und Schülern zu differenzieren.

4.4.5 Methodenkritik

Die Items wurden ursprünglich sowohl für Schülerinnen und Schüler des Gymnasiums als auch der Realschule entwickelt (vgl. Kapitel 2.1.2), waren für die inzwischen eingeführte Sekundarschule jedoch nicht geeignet (vgl. Kapitel 3.3.4). Die Zielgruppe musste daher eingeschränkt werden, so dass der Itempool für die Schülerinnen und Schüler des Gymnasiums eher leicht und die Stichprobe homogener als ursprünglich intendiert war. Die Ergebnisse sind deshalb mit diesen Einschränkungen zu betrachten. Um mit Blick auf die Bildungsstandards (KMK, 2005) die Modellkompetenz von Schülerinnen und Schülern allgemeinbildender Schulen zu erheben, müsste das Instrument entsprechend angepasst werden.

Für die Schülerinnen und Schüler des Gymnasiums zeigte die Methode des lauten Denkens, dass sich ihre kognitiven Prozesse bei der Itembearbeitung auf Modellkompetenz beziehen und die Itembearbeitung somit als Indikator für Modellkompetenz interpretierbar ist (vgl. Kapitel 3.3.4). Dass die Items das Kompetenzmodell angemessen repräsentieren, wurde bereits während der Testentwicklung durch Expertenratings der Konstruktionsanleitung sowie der einzelnen Items geklärt (vgl. Kapitel 3.1.4, 3.2). Insgesamt wurde über die systematische Instrumentenentwicklung ein enger Bezug zwischen der theoretischen Grundlage und der Operationalisierung in MC-Items hergestellt. Aus diesem Grund war der Test selbst vermutlich nicht ausschlaggebend für die – auch für eine solche eher homogene Stichprobe – geringe Varianz, die wie die damit verbundene geringe Reliabilität die Interpretation der hier vorgestellten Befunde erwschwert. Als Gründe für die geringe Varianz kommen damit die Beschreibung des Konstrukts sowie die Gestaltung des Unterrichts mit Blick auf Modellkompetenz in Frage. Der Vergleich der Strukturmodelle bietet wenig belastbare Argumente für die theoriegeleiteten Strukturen und trägt entsprechend nicht zur Klärung der Ursachen für die geringe Varianz bei. Die Unterscheidung der Niveaus, Unterschiede zwischen den Jahrgangsstufen sowie die Befunde zur diskriminanten Validität können schon eher als Hinweise darauf interpretiert werden, dass die geringe Varianz darauf zurückzuführen ist, dass Schülerinnen und Schüler wenig Gelegenheiten zum Kompetenzerwerb haben.

Dies kann auf Grundlage der hier diskutierten Daten jedoch nicht mit Sicherheit festgestellt werden.

Die Daten wurden in ganzen Klassen an acht Schulen erhoben, so dass eine proportional geschichtete Klumpenstichprobe vorliegt (Bortz, 2005). Weil davon ausgegangen werden muss, dass die Lerngelegenheiten die Modellkompetenz der Schülerinnen und Schüler beeinflussen, kann die Zugehörigkeit zu einer Klasse bzw. Schule die Ergebnisse verzerren. Effekte von Klassen- bzw. Schulvariablen auf Individualvariablen könnte man in weiteren Untersuchungen mit Strukturgleichungsmodellen auf mehreren Ebenen modellieren. Indem in diesem Rahmen Lerngelegenheiten zusätzlich erhoben werden, kann auch abschließend geklärt werden, ob diese tatsächlich die geringe Varianz begründen.

5 Modellkompetenz im Kontext Biologieunterricht

Neben Erkenntnissen, die im Bereich der Grundlagenforschung (für Kernergebnisse siehe Kapitel 6) und für Folgeprojekte (Kapitel 7) relevant sind, führte das hier berichtete Projekt auch zu Erkenntnissen, die als Anknüpfungspunkte für die Förderung von Modellkompetenz genutzt werden können (Abb. 24).

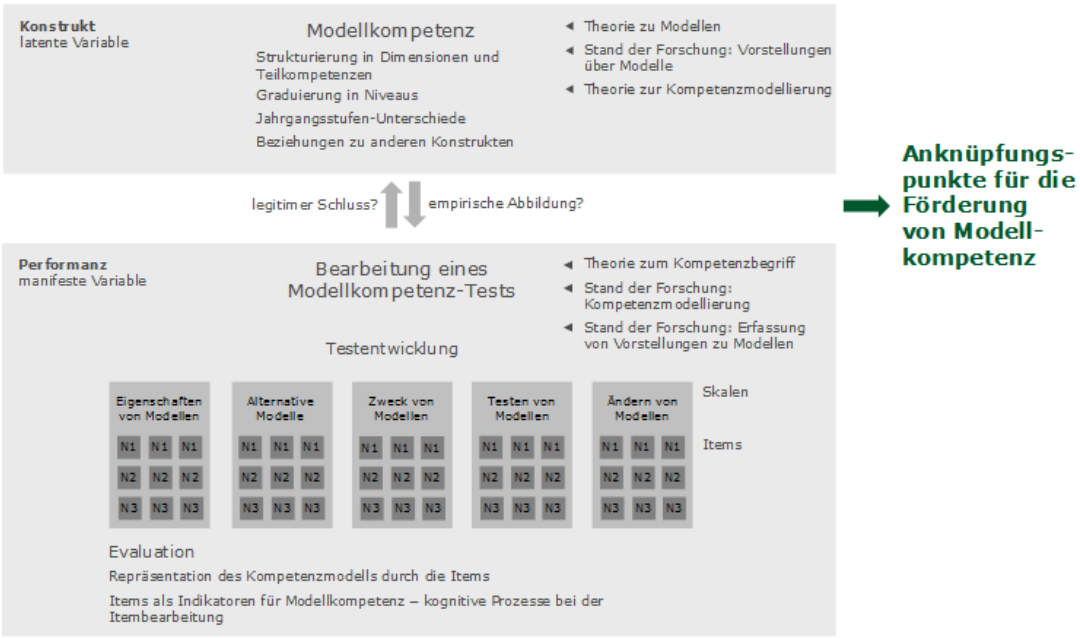


Abb. 24: Konzeption des Projekts – Anknüpfungspunkte für die Förderung von Modellkompetenz.

Ganz grundsätzlich deuten die Unterschiede zwischen den Jahrgangsstufen sowie die Abgrenzbarkeit von allgemeinen kognitiven Fähigkeiten darauf hin, dass Modellkompetenz erlern- und vermittelbar ist (vgl. Kapitel 4.4.3, 4.4.4). Damit ist sie der unterrichtlichen Förderung zugänglich. Da empirisch eine mediale und eine methodische Perspektive, je Teilkompetenz differenziert in drei Niveaus, voneinander abgrenzbar sind (vgl. Kapitel 4.4.2), können sie im Biologieunterricht gezielt herangezogen werden (vgl. Henze et al., 2007). Zum

einen kann darüber gezielt der Anspruch des Unterrichts an die Lernvoraussetzungen der Schülerinnen und Schüler angepasst werden. Zum anderen können über die Perspektiven auf Modelle unterschiedliche Ziele verfolgt werden (vgl. Kapitel 2.2.4): Um das Ziel *learning science* (Hodson, 1993) zu erreichen und den Erwerb von Fachwissen zu fördern, können Modelle unter einer medialen Perspektive eingesetzt werden. Diese betrifft die Beurteilung der Passung zwischen Theorie, Empirie und Modell und damit die Theoriebildung (vgl. Kapitel 2.2.2). Unter einer methodischen Perspektive auf Modelle, die deren Funktion im Erkenntnisprozess thematisiert (vgl. Kapitel 2.2.3), können die Ziele *learning about science* und *learning how to do science* (Hodson, 1993) erreicht werden. Dabei wird die Gültigkeit von Theorien beurteilt, indem Hypothesen über die Passung von Empirie und Theorie mithilfe von Modellen geprüft werden. Als Lernvoraussetzungen können Lehrerinnen und Lehrer bis Jahrgangsstufe 7 bei Schülerinnen und Schülern vor allem mit Kompetenzen im medialen, dann zunehmend im methodischen Bereich rechnen (vgl. Kapitel 4.4.3).

Durch die hier entwickelten Testaufgaben wird Modellkompetenz konkretisiert und enger an den Unterricht angebunden (vgl. Bernholt et al., 2009; Klieme, Avenarius et al., 2007). Für den Unterricht können sie als Grundlage zur Entwicklung von Lernaufgaben genutzt werden. Dabei sollte beachtet werden, dass das mentale Modell, das einem konzeptuellen Modell zugrunde liegt, nicht mit dem mentalen Modell überstimmen muss, das Schülerinnen und Schüler zum konzeptuellen Modell entwickeln (Greca & Moreira, 2000; Justi & Gilbert, 2003; Norman, 1983). Lehrerinnen und Lehrer sollten deshalb ggf. im Unterricht sicherstellen, dass Schülerinnen und Schüler ein vorgegebenes Modell so verstehen, wie es intendiert ist.

Beim konkreten Umgang mit Modellen scheinen Schülerinnen und Schüler entsprechend der Einschätzung von Schwarz et al. (2009) auf Vorstellungen aus verschiedenen Teilkompetenzen zurückzugreifen. Hierfür sprechen sowohl die Befunde zu kognitiven Prozessen bei der Itembearbeitung als auch Hinweise auf Zusammenhänge zwischen den Teilkompetenzen bei der Strukturierung von Modellkompetenz (vgl. Kapitel 3.3.4, 4.4.1). Wenn bei den Schüle-

rinnen und Schülern ein vernetztes Wissen gefördert werden soll, bietet es sich an, solche Bezüge zwischen verschiedenen Teilkompetenzen aufzugreifen und Unterrichtsangebote entsprechend zu gestalten. Auch wenn es Hinweise darauf gibt, dass Modellkompetenz im Sinne eines konkreten Umgangs mit Modellen eine globale Fähigkeit darstellt, kann die Strukturierung in Teilkompetenzen als inhaltliche Bereiche für die Förderung genutzt werden.

Eine zentrale Rolle scheint der Vergleich von Modell und Original zu spielen, der sowohl in den Experten-Ratings als auch in den Denkprotokollen der Schülerinnen und Schüler häufig deutlich wurde (vgl. Kapitel 3.1.4, 3.2, 3.3.4). Diese grundlegende Fähigkeit sollte an geeigneten Beispielen geschult und explizit thematisiert werden, da inadäquate Parallelisierungen möglicherweise das gesamte Verständnis von Modellen behindern können. Außerdem sollte darauf geachtet werden, dass dennoch die Intentionalität von Modellen deutlich wird (Giere, 2009; Stachowiak, 1973, 1983). Die Intentionalität kann neben dem unterschiedlichen Abstraktionsgrad von Theorie und Modell sowie der Vermittlung des Modells zwischen Theorie und Empirie als Ansatzpunkt zur Reflexion dienen.

Einige Kategorien zu Vorstellungen zu Modellen (Kapitel 2.3.2) sind eng miteinander verknüpft: Grosslight et al. (1991) bringen z. B. ein einfaches Modellverständnis (‚Eigenschaften von Modellen', Niveau I bis II) mit einer medialen Sicht auf ‚Alternative Modelle' (Niveau I bis II) in Verbindung, ein komplexes (‚Eigenschaften von Modellen', Niveau III) mit einer methodischen Sicht (‚Alternative Modelle', Niveau III). Die Befunde zu kognitiven Prozessen bei der Itembearbeitung legen nahe, dass die Fähigkeit, die Passung zwischen Modell und Original zu beurteilen (‚Testen von Modellen', Niveau II), für die Änderung eines Modells *von* etwas (Mahr, 2008a; ‚Ändern von Modellen', Niveau II) benötigt wird (vgl. Kapitel 3.3.4).

Außerdem deutet sich in den Denkprotokollen der Schülerinnen und Schüler an, dass die Fähigkeiten, Modelle experimentell zu testen (‚Testen von Modellen', Niveau III), die Fähigkeiten zur Änderung von Modellen auf der Grundlage experimentell gewonnener Daten (‚Ändern von Modellen', Niveau III) be-

einflussen (vgl. Kapitel 3.3.4). Für diesen Teilbereich von Modellkompetenz spielt außerdem methodisches Wissen über Ziele und Vorgehensweisen beim Experimentieren und damit das wissenschaftliche Denken und Wissenschaftsverständnis nach Mayer (2007) eine Rolle. Wenn Schülerinnen und Schüler beim ‚Testen' und insbesondere beim ‚Ändern von Modellen' Schwierigkeiten haben, ist ein möglicher Ansatzpunkt zur Förderung entsprechend auch die Förderung von Kompetenzen in den Bereichen wissenschaftliches Denken und Wissenschaftsverständnis. Dabei könnten inhaltliche Eingangsüberzeugungen und Strategien thematisiert werden (Chinn & Brewer, 1993; Hammann et al., 2006).

Darüber hinaus sprechen sowohl die Befunde zu kognitiven Prozessen von Schülerinnen und Schülern bei der Itembearbeitung als auch unterschiedliche Befunde zu unterschiedlich abstrakten, domänen- und kontextspezifischen Aufgabenformaten dafür, wie Leisner-Bodenthin (2006) sowie Leisner und Mikelskis (2004) domänenspezifische und -übergreifende Modellkompetenz zu unterscheiden. Damit ist es möglicherweise sinnvoll, im Biologieunterricht gezielt Phasen mit einer abstrakten Thematisierung von Modellen mit Phasen der Reflexion über Modelle in spezifischen Kontexten abzuwechseln. Epistemologische Vorstellungen über Modelle können evtl. angebunden an einen Kontext, aber nicht als abstrakter Standpunkt formuliert werden (Sins et al., 2009). Wenn dies bei Schülerinnen und Schülern beobachtet wird und ein abstraktes Verständnis gefördert werden soll, könnte es hilfreich sein, von konkreten Kontexten auszugehen. Beim ‚Ändern von Modellen' fällt es Schülerinnen und Schülern z. B. offenbar leichter, die Veränderbarkeit von konkreten Modellen zu begründen, als abstrakte Positionen hierzu zu formulieren (AAAS, o. J.; Grosslight et al., 1991; Trier & Upmeier zu Belzen, 2009). Umgekehrt wird in den kognitiven Prozessen bei einigen Schülerinnen und Schülern deutlich, dass sie über ein abstraktes Verständnis verfügen, ohne dass sie es am konkreten Modell umsetzen können. Auch Leisner und Mikelskis (2004) beobachteten einen Zuwachs im domänenübergreifenden Wissen, aber ein unklares Bild für das domänenspezifische Wissen. Entsprechend sollten ggf. spezifische Kontexte gezielt aus Ausgangspunkte für die Reflexion über Modelle genutzt werden.

6 Fazit

Die Ergebnisse der vorliegenden Untersuchung legen für den hier untersuchten konkreten Umgang mit Modellen empirisch eine eindimensionale Struktur von Modellkompetenz nahe. Gleichzeitig gibt es Indizien für spezifische Verknüpfungen zwischen den Teilkompetenzen, z. B. zwischen dem ‚Testen' und ‚Ändern von Modellen', sowie für den Vergleich von Modell und Original als zentrales Konzept. Diese Hinweise sollten in anschließenden Studien weiterverfolgt werden, da die Datenbasis der vorliegenden Untersuchung keine abschließenden Einschätzungen zulässt.

Mit Blick auf die Graduierung von Modellkompetenz zeigte sich insgesamt, dass sich empirisch drei Niveaus mit ansteigenden Anforderungen an die Schülerinnen und Schüler abbilden. Insbesondere in der Teilkompetenz ‚Testen von Modellen' wäre jedoch die Formulierung der Niveaus zu überdenken.

Der Unterschied in der Performanz im Bereich Modellkompetenz war über die Jahrgangsstufen 7 bis 10 insgesamt nicht sehr groß, verschiebt aber die Kompetenzen der Schülerinnen und Schüler von einem medial geprägten in einen zunehmend methodischen Bereich. Die Unterschiede bestanden vor allem zwischen den Jahrgangsstufen 8 und 9. Worauf dies zurückzuführen ist, könnte eine Betrachtung der Lerngelegenheiten zeigen. Das Alter spielte für den Kompetenzerwerb keine nennenswerte Rolle.

Sowohl Hinweise aus der querschnittlichen Betrachtung von Modellkompetenz als auch ihre Beziehung zu allgemeinen kognitiven Fähigkeiten sprechen dafür, dass Modellkompetenz erlernbar ist. Gegen allgemeine kognitive Fähigkeiten und Lesefähigkeiten kann sie im Sinne einer diskriminanten Validierung sinnvoll abgegrenzt werden. Zur Domänenspezifität und damit zur konvergenten Validierung zeichnen die Befunde jedoch ein unklares Bild, da Modellkompetenz mit Leistungen in naturwissenschaftlichen und sprachlichen Fächern in ähnlicher Höhe zusammenhing. Zwischen dem Wissenschaftsverständnis und Modellkompetenz bestand, anders als theoriegeleitet anzunehmen, kein Zu-

sammenhang. Zwischen Schülerinnen und Schülern waren keine Unterschiede zu beobachten.

Insgesamt deuten die Befunde dieser Studie darauf hin, dass kontextspezifische und -unabhängige Facetten von Modellkompetenz und somit der konkrete Umgang mit Modellen und ein abstraktes Modellverständnis unterschieden werden können.

7 Ausblick

Über dieses Projekt hinweg ergaben sich Hinweise darauf, dass ein Unterschied zwischen domänen- bzw. kontextunabhängigen und domänen- bzw. kontextspezifischen Aufgaben besteht. Eine interessante anschließende Fragestellung ist deshalb die nach dem Zusammenhang der verschiedenen, unterschiedlich stark domänen- und kontextspezifischen Antwortformate zu Modellkompetenz (MC- sowie FC-Items und Items in offenem Antwortformat). Die Höhe von Korrelationen zwischen ihnen kann Aufschluss darüber geben, inwiefern die Antwortformate das Gleiche messen. Um zu klären, welche Merkmale von Kontexten für die Kontextspezifität relevant sind, ist zusätzlich eine systematische Variierung von Itemkontexten notwendig.

Die Frage nach der Domänen- und Kontextabhängigkeit von Modellkompetenz berührt auch die Frage, inwiefern explizites Wissen für den Umgang mit Modellen notwendig ist und ob es über die implizite Anwendung von Modellen hinaus relevante Kompetenzen darstellt (vgl. Schwarz & White, 2005). Auch die Übertragbarkeit von Befunden für die Domäne Biologie auf andere Naturwissenschaften sollte in diesem Zusammenhang diskutiert werden.

Darüber hinaus erscheinen aus mehreren Gründen anschließende Untersuchungen sinnvoll, die den vorliegenden MC-Itempool erweitern. Zum einen sollten Aufgaben für andere Zielgruppen, z. B. Schülerinnen und Schüler der Sekundarschule, Studierende oder Lehrerinnen und Lehrer, adaptiert bzw. entwickelt werden, um die Modellkompetenz dieser Personengruppen erheben zu können. Dies würde zum einen ermöglichen, Aussagen über alle Schülerinnen und Schüler an allgemeinbildenden Schulen zu treffen. Zum anderen könnte geprüft werden, welche Kompetenzausprägung auf Seiten der (angehenden und erfahrenen) Lehrerinnen und Lehrer vorliegt, um einen Einfluss auf die Modellkompetenz der Schülerinnen und Schüler prüfen zu können. Mit dem Projekt KoWADIS (Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Arbeits- und Denkweisen bei Studierenden (Lehramt) in den drei naturwissenschaftlichen Fächern Biologie, Chemie und Physik) wird u. a. diese Fragestellung bei Studierenden verfolgt

(Hartmann, Upmeier zu Belzen & Krüger, 2012). Auch die Diagnosekompetenz von Lehrerinnen und Lehrern für die Modellkompetenz von Schülerinnen und Schülern ist eine relevante Variable für weitere Untersuchungen (vgl. Dübbelde et al., 2011).

Ein weiterer Grund, zusätzliche Aufgaben zu entwickeln, bezieht sich auf ihren möglichen Verwendungszweck. Der hier entwickelte Itempool ist inhaltlich breit angelegt und zielt nicht auf eine große Messgenauigkeit auf Individualebene. Für Individualdiagnosen müssten entsprechend zusätzliche Items entwickelt werden, die in einzelnen inhaltlichen Bereichen stärker in die Tiefe gehen (vgl. Leutner et al., 2007) und ggf. sogar an spezifische Unterrichtsinhalte angebunden sind. Auch für diesen Anwendungskontext könnte es sinnvoll sein, Lehrerinnen und Lehrer an der Aufgabenentwicklung zu beteiligen, so dass eine breite Themenvielfalt, Praxisnähe und Eignung der Items für die Zielgruppe sichergestellt werden. Die Entwicklung weiterer Aufgaben kann dabei auf die hier entwickelte Konstruktionsanleitung (Kapitel 3.1.3) sowie die hier formulierten Anforderungen an Modelle (Kapitel 3.1.5) aufbauen. Aufgabenkontexte, die physikalische und/oder chemische Bezüge herstellen, sollten ggf. gezielt zur Variierung der Schwierigkeit eingesetzt werden (vgl. Kapitel 3.1.6). Bei der Entscheidung, wie viele Items der jeweiligen Zielgruppe vorgelegt werden, sollte in Vorstudien geprüft werden, inwiefern ein *Optimizing-Satisficing*-Problem (Jonkisz et al., 2012) auftritt und wie viele Items sinnvoll bearbeitet werden können. Bei der Entwicklung von weiteren Items wäre es sinnvoll, neben den Niveaus weitere schwierigkeitserzeugende Merkmale gezielt zu variieren und zu untersuchen, um die Itembearbeitung besser zu verstehen.

Mit einem erweiterten Itempool können über die Untersuchung weiterer Zielgruppen und die Verfolgung weiterer Anwendungskontexte hinaus die Ergebnisse des vorliegenden Projekts differenziert werden. Wie Vorstellungen in einzelnen Niveaus miteinander zusammenhängen und inwiefern sich eine Person je nach Teilkompetenz in unterschiedlichen Niveaus bewegt (vgl. Justi & Gilbert, 2003), ist eine lohnenswerte Fragestellung für weitere Projekte. Neben der Modellierung von Zusammenhängen zwischen einzelnen Niveaus von

Teilkompetenzen könnte hierfür über Items in offenem Antwortformat das Kategoriensystem von Grünkorn et al. (in Vorb.) herangezogen werden, das die Niveaus je Teilkompetenz in Unterkategorien differenziert. Solche Studien könnten z. B. die Reihung der Niveaus in einzelnen Teilkompetenzen prüfen. Auch die hier formulierten detaillierten Hypothesen zu Bezügen zwischen Teilkompetenzen und damit zur Struktur von Modellkompetenz (**H13**, **H14**, **H15a/b**) könnten in Folgestudien geprüft werden. Indizien, die in der vorliegenden Untersuchung für die Annahme von **H13** und **H14** und damit für Bezüge zwischen den Teilkompetenzen ‚Testen' und ‚Ändern von Modellen' sprechen, könnten weiterverfolgt werden. Während die Beziehung zwischen dem ‚Zweck von Modellen', Niveau I, und den ‚Kenntnissen über Modelle' (**H15a**) mit den vorliegenden Daten nicht zu prüfen war, erscheint eine Einordnung dieser Teilkompetenz in die ‚Kenntnisse über Modelle' (**H15b**) auf Grundlage der quantitativen Daten eher unwahrscheinlich. Auch die Rolle des Vergleichs von Modell und Original und damit die Rolle der Teilkompetenz ‚Eigenschaften von Modellen' (z. B. der Zusammenhang zum ‚Testen von Modellen', Niveau II, (**H11**) oder zum ‚Zweck von Modellen', Niveau I, (**H12**)) wäre ein lohnenswerter Fokus für weitere Untersuchungen. Eine solche explizite Modellierung von Bezügen zwischen den Teilkompetenzen, auch in einzelnen Niveaus, wäre über Strukturgleichungsmodelle möglich. Hierfür müsste ein Testheftdesign erstellt werden, das mehr Items je Person oder weniger Bereiche von Modellkompetenz vorgibt und damit eine Abdeckung der Kovarianzmatrix von mindestens 0.1 aufweist (Muthén & Muthén, 2007). Auf dieser Grundlage könnten Zusammenhängen innerhalb von Modellkompetenz untersucht werden, die differenzierte Hinweise für die Gestaltung von Unterrichtsangeboten bieten.

Zusätzliche Erklärungen zu hier erzielten Befunden könnte außerdem die Erhebung der Lerngelegenheiten von Schülerinnen und Schülern bieten. Dies betrifft zum einen die Erklärung der geringen Varianz durch die Implementierung von Modellkompetenz in den Unterricht sowie zum anderen weitere Hinweise zur Kompetenzentwicklung. Inwiefern das Kompetenzmodell auch ein Kompetenzentwicklungsmodell darstellt, muss anhand intraindividueller Unterschiede in einem längsschnittlichen Design geprüft werden. Dieser Fragestellung gehen Patzke und Upmeier zu Belzen (2011, 2012) nach und verwen-

den dort Lerngelegenheiten als erklärende Variable für die Kompetenzentwicklung.

Indem in einem weiteren Anschlussprojekt Modellkompetenz und Kompetenzen im wissenschaftlichen Denken erhoben werden, könnte geklärt werden, inwiefern die Kontextabhängigkeit der epistemologischen Vorstellungen von Schülerinnen und Schülern, z. B. beim ‚Ändern von Modellen', Niveau III, relevant ist. Auch Bezüge zur *representational competence*, Diagrammkompetenz sowie Systemkompetenz können sinnvolle Anknüpfungspunkte für eine gezielte Förderung dieser Kompetenzen wie Modellkompetenz bieten. Die Rolle des Chemieunterrichts für Modellkompetenz kann in Projekten geklärt werden, die beide Fächer betrachten (z. B. VerE; Nowak et al., 2012).

Im hier vorgestellten Projekt ging es zunächst um Grundlagenforschung und nicht darum, die Items in der Schule einzusetzen, um Lernvoraussetzungen zu diagnostizieren. Der Forschungsbereich der Nutzung der diagnostischen Information in der Schule nach Klieme et al. (2008) bleibt demnach noch offen.

Erste Ansätze zur Gestaltung von Unterrichtsangeboten, die den Erwerb von Modellkompetenz fördern, arbeiten Fleige, Seegers, Upmeier zu Belzen und Krüger (2012a, 2012b) heraus und stellen Materialien für den Unterricht bereit. Orsenne und Upmeier zu Belzen (2012) untersuchen den Einfluss verschiedener Reflexionsangebote auf Vorstellungen beim Umgang mit Modellen. Aufbauend auf diese beiden Studien sowie Befunde der vorliegenden Arbeit können Fördermodule entwickelt, evaluiert und Lehrerinnen und Lehrern zur Verfügung gestellt werden.

Dank

Dieses Projekt hätte ohne die Beteiligung vieler anderer so nicht abgeschlossen werden können. Mein Dank gilt den Schülerinnen und Schülern sowie den Kollegien der Schulen, die dieses Projekt unterstützt haben und bereit waren, an Befragungen teilzunehmen. Dem Bundesministerium für Bildung und Forschung danke ich für die Förderung im Rahmen des Nachwuchsprogramms für die empirische Bildungsforschung seit 2009, das mir Unterstützung durch Expertenberatungen, die Teilnahme an Workshops und fruchtbare Diskussionen mit Kolleginnen und Kollegen aus verschiedenen Disziplinen ermöglicht hat.

Für die Begleitung dieser Arbeit gebührt vor allem meiner Doktormutter Prof. Dr. Annette Upmeier zu Belzen Dank. Sie hat sowohl mir als auch meinem Projekt durch ihren Anspruch, zunächst Klarheit über theoretische Fundierung, Ziele und Fragestellungen zu gewinnen und weitere Entscheidungen systematisch darauf zu stützen, Wurzeln gegeben. Außerdem hat sie darauf bestanden, dass manche Dinge sich entwickeln müssen und man dafür bei aller Zielgerichtetheit offen bleiben muss. Diese Haltung nehme ich nicht nur für wissenschaftliche Kontexte mit.

Prof. Dr. Olaf Köller half mir von Beginn an, einen Weg durch die Vielfalt der Methoden zu finden. Dabei gab er mir immer die Zuversicht, auch diese Herausforderungen zu meistern. Darin haben mich ebenfalls Prof. Dr. Andreas Frey und insbesondere Prof. Dr. Johannes Hartig unterstützt, denen ich für ihren Rat und ihre Bereitschaft zum Austausch danke.

Unser Kooperationspartner zum Thema Modellkompetenz, Prof. Dr. Dirk Krüger, hat über den Verlauf des Projekts kritisch meine Entscheidungen hinterfragt und somit zu deren Reflektiertheit beigetragen. Auch die Diskussionen mit den anderen Doktorandinnen und Doktoranden, die zum Thema Modelle im Biologieunterricht arbeiten, haben mir dabei geholfen.

Ein Dankeschön geht an meine Kolleginnen und Kollegen aus der Arbeitsgruppe der HU, die mir während meiner Promotion mit Rat und Tat zur Seite stan-

den – sei es mit einer kreativen Kaffeepause, in einer schwierigen Phase oder bei der Durchführung meiner Datenerhebungen. Diese Unterstützung fand nicht nur innerhalb von Berlin statt, sondern durch Ulrike Trier zeitweise sogar international. Ralf Merkel hat mir dankenswerterweise beim Bau der Modelle für die Aufgaben geholfen und früh morgens Berge an Material zur Schule transportiert. Juliane Orsenne danke ich für ihr fachkundiges Urteil über die sprachliche Komplexität meiner Aufgaben sowie ihre kreativen Ideen für Aufgabenkontexte. Insbesondere an Lösungen zu methodischen Problemen war Stefan Hartmann mit seinem Rat und Literaturtipps beteiligt. Eine unschätzbar wertvolle Hilfe waren Nadine Szymanowski, wegen der ich an viele Dinge nicht mehr selbst denken musste, weil ich sie bei ihr in guten Händen wusste, und ihre würdige Vertretung Jenny Voigtländer.

Viele Überlegungen wurden im Austausch mit anderen Doktorandinnen von einer groben Idee zu einem formulierbaren Gedanken ausgeschärft. Der Austausch mit Dr. Nicole Wellnitz hat mich insbesondere in der Endphase meiner Promotion sehr weitergebracht. Dr. Cornelia Sander, Sarah Huch und Sarah Dannemann haben mein Projekt über seine gesamte Dauer kritisch begleitet, Schwierigkeiten aufgezeigt und gleichzeitig Lösungen vorgeschlagen. Nicht nur deshalb freue ich mich, euch kennengelernt zu haben! Juliane Grünkorn, Christiane Patzke und Dr. Sandra Nitz möchte ich besonders für den Gedankenaustausch danken – er war nicht nur konstruktiv und gewinnbringend, sondern immer eine große Freude. Es war eine großartige Zeit mit euch.

Außerdem danke ich meiner Familie und meinen Freunden, insbesondere Jonas Hurlin und Dr. Anna von Hopffgarten, für die Unterstützung während der Promotionszeit. Sie haben mich zuversichtlich und gelassen durch die Höhen und Tiefen begleitet.

Literaturverzeichnis

AAAS (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

AAAS. (o. J.). AAAS Science Assessment beta: Topic Models. Zugriff am 01.10.2012. Verfügbar unter http://assessment.aaas.org/topics/MO#/

Abel, G. (2008). Modell und Wirklichkeit. In U. Dirks & E. Knobloch (Hrsg.), *Modelle* (S. 31–45). Frankfurt a. M.: Peter Lang.

Adams, R. J., Wilson, M. & Wang, W. (1997). The Multidimensional Random Coefficients Multinominal Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23.

Aikenhead, G. & Ryan, A. (1992). The development of a new instrument: 'Views on Science-Technology-Society' (VOSTS). *Science Education*, *76*, 477–492.

Ainsworth, S. (2008). The Educational Value of Multiple-Representationa when Learning Complex Scientific Concepts. In J. K. Gilbert, M. Reiner & M. Nakhleh (Hrsg.), *Visualization. Theory and Practice in Science Education* (S. 191–208). Dordrecht: Springer.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage University Paper No. 136.

Amelang, M. & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention*. Berlin: Springer.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2006). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Berlin: Springer.

Baddeley, A. D. (2002). Is Working Memory Still Working? *European Psychologist*, *7*(2), 85–97.

Bailer-Jones, D. M. (1999). Tracing the Development of Models in the Philosophy of Science. In L. Magnani, N. J. Nersessian & P. Thagard (Hrsg.), *Model-based reasoning in scientific discovery* (S. 23–40). New York: Kluwer Academic/Plenum Publishers.

Bailer-Jones, D. M. (2000). *Naturwissenschaftliche Modelle: Von Epistemonologie zu Ontologie*. Bielefeld: Kongress der Gesellschaft für Analytische Philosophie GAP 4.

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedien: Erfassung, Beschreibung und Vermittlung wirksamer metakognitiver Strategien und Regulationsaktivitäten*. Münster: Waxmann.

Baumert, J. (1997). Scientific Literacy: A German Perspective. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy* (S. 167–180). Kiel: Institute for Science Education (IPN).

Bayrhuber, H., Bögeholz, S., Eggert, S., Elster, D., Grube, C., Hößle, C. et al. (2007). Biologie im Kontext: Erste Forschungsergebnisse. *MNU*, *60*(5), 304–313.

Bernholt, S., Parchmann, I. & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *ZfdN*, *15*, 219–245.

Bilandzic, H. (2005). Lautes Denken. In L. Mikos & C. Wegener (Hrsg.), *Qualitative Medienforschung. Ein Handbuch* (S. 362–370). Konstanz: UVK.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Wien: Springer.

Bortz, J. & Döring, N. (2005). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer.

Brandstädter, K., Harms, U. & Großschedl, J. (2012). Assessing System Thinking Through Different Concept-Mapping Practices. *International Journal of Science Education*, 34(14), 2147-2170.

Buckley, B. C., & Boulter, C. J. (2000). Investigating the Role of Representations and Expressed Models in Building Mental ModelS. In J. K. Gilbert & C. J. Boulter (Hrsg.), *Developing Models in Science Education* (S. 119–135). Dordrecht: Kluwer.

Bybee, R. W. (1997). Toward an Understanding of Scientific Literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy* (S. 37–68). Kiel: Institute for Science Education (IPN).

Bybee, R. W. (2002). Scientific Literacy - Mythos oder Realität? In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific Literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (S. 21–43). Opladen: Leske & Budrich.

Campbell, N. A., Kratochwil, A., Lazar, T. & Reece, J. B. (2009). *Biologie*. München: Pearson Studium.

Carey, S., Evans, R., Honda, M., Jay, E. & Unger, C. (1989). "An experiment is when you try it and see if it works": a study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*(Special Issue), 514–529.

Carnap, R. (1928). Der logische Aufbau der Welt. Berlin: Weltkreis.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Carstensen, C. H., Frey, A., Walter, O. & Knoll, S. (2007). Technische Grundlagen des dritten internationalen Vergleichs. In Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg). *PISA 2006. Die Ergebnissse der dritten internationalen Vergleichsstudie* (S. 367-390). Münster: Waxmann.

Chinn, C. A. & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49.

Chittleborough, G. & Treagust, D. F. (2007). The modelling ability of non-major chemistry students and their understanding of the sub-microscopic level. *Chemistry Education Research and Practice*, *8*(3), 274–292.

Chomsky, N. (1968). *Language and mind*. New York: Harcourt, Brace & World.

Colbourn, C. J. & Dinitz, J. H. (1996). *The CRC Handbook of Combinatorial Designs.* Boca Raton, FL: CRC Press.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Connell, M. W., Sheridan, K. & Gardner, H. (2003). On Abilites and Domains. In R. J. Sternberg & E. Grigorenko (Hrsg.), *The Psychology of Abilities, Competencies, and Expertise* (S. 126–155). Cambridge, NY: Cambridge University Press.

Crawford, B. & Cullin, M. (2005). Dynamic assessments of preservice teachers' knowledge of models and modelling. In K. Boersma, H. Eijkelhof, M. Goedhart & O. Jong (Hrsg.), *Research and the Quality of Science Education* (S. 309–323). Dordrecht: Springer.

da Costa, N. & French, S. (2000). Models, Theories, and Structures: Thirty Years on. *Philosophy of Science*, (Proceedings), 116–127.

Develaki, M. (2007). The Model-Based View of Scientific Theories and the Structuring of School Science Programmes. *Science & Education*, *16*, 725–749.

Devitt, M. (2005). Scientific Realism. In F. Jackson & M. Smith (Hrsg.), *The Oxford handbook of contemporary philosophy* (S. 767–791). Oxford: Oxford University Press.

Driver, R., Leach, J., Millar, R. & Scott, P. (1996). *Young people's images of science*. Buckingham: Open University Press.

Dübbelde, G., Mayer, J., Möller, A. & Aufschnaiter, C. von. (2011). Diagnostische Kompetenzen angehender Biologielehrer [Abstract]. In F. X. Bogner (Hrsg.), *Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIO*. (S. 112–113). Bayreuth: Difo Druck.

Duit, R. (1991). On the Role of Analogies and Metaphors in Learning Science. *Science Education*, *75*(6), 649–672.

Eggert, S. (2008). *Bewertungskompetenz für den Biologieunterricht - Vom Modell zur empirischen Überprüfung*. Dissertation. Zugriff am 01.10.2012. Verfügbar unter http://webdoc.sub.gwdg.de/diss/2008/eggert/eggert.pdf

Einhaus, E. & Schecker, H. (2006). Item-Merkmale im Expertenrating. In A. Pitton (Hrsg.), *Lehren und Lernen mit neuen Medien* (S. 111–113). Berlin: LiT.

Embretson, S. E. (1983). Construct Validity: Construct Representation Versus Nomothetic Span. *Psychological Bulletin*, *93*(1), 179–197.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Ericsson, K. A. & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review*, *87*(3), 215–251.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fleige, J., Seegers, A., Upmeier zu Belzen, A. & Krüger, D. (2012a). *Modellkompetenz im Biologieunterricht Klasse 7 - 10*. Donauwörth: Auer.

Fleige, J., Seegers, A., Upmeier zu Belzen, A. & Krüger, D. (2012b). Förderung von Modellkompetenz im Biologieunterricht. *MNU*, *65*(1).

Frey, A. J., Hartig, J. & Rupp, A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement*, *28*(3), 39–53.

Frigg, R. (2006). Scientific Representation and the Semantic View of Theories. *Theoria*, *55*, 49–65.

Frigg, R. & Hartmann, S. (2012). Models in science. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Zugriff am 01.10.2012. Verfügbar unter http://plato.stanford.edu/archives/fall2012/entries/models-science/

Gentner, D. (1989). The mechanism of analogical learning. In S. Vosniadou (Hrsg.), *Similarity and analogical reasoning* (S. 199–241). Cambridge, NY: Cambridge University Press.

Gentner, D. (2002). Analogy in Scientific Discovery: The Case of Johannes Kepler. In L. Magnani & N. J. Nersessian (Hrsg.), *Model-based reasoning. Science, technology, values* (S. 21–40). New York: Kluwer Academic.

Giere, R. N. (1988). *Explaining science: A cognitive approach. Science and its conceptual foundations.* Chicago, IL: University of Chicago Press.

Giere, R. N. (1999). *Science without laws. Science and its conceptual foundations.* Chicago, IL: University of Chicago Press.

Giere, R. N. (2001). A New Framework for Teaching Scientific Reasoning. *Argumentation*, *15*, 21–33.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *73*, 742–752.

Giere, R. N. (2006). *Scientific perspectivism*. Chicago, IL: University of Chicago Press.

Giere, R. N. (2009). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Giere, R. N., Bickle, J. & Mauldin, R. F. (2006). *Understanding scientific reasoning*. Belmont, CA: Thomson Wadsworth.

Giesbrecht, F. G. & Gumpertz, M. L. (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments.* Hoboken: Wiley.

Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching*, *28*(1), 73–79.

Gobert, J. D. & Pallant, A. (2004). Fostering Students' Epistemologies of Models via Authentic Model-Based Tasks. *Journal of Science Education and Technology*, *13*(1), 7–22.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576.

Graham, J. W., Cumsille, P. E. & Elek-Fisk, E. (2003). Methods for Handling Missing Data. In J. A. Schinker & W. F. Velicer (Hrsg.), *Handbook of psychology* (S. 87–114). New York: Wiley.

Greca, I. M. & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education*, *22*(1), 1–11.

Gropengießer, H. (2001). *Didaktische Rekonstruktion des Sehens: Wissenschaftliche Theorien und die Sicht der Schüler in der Perspektive der Vermittlung*. Oldenburg: Didaktisches Zentrum.

Grosslight, L., Unger, C., Jay, E. & Smith, C. L. (1991). Understanding Models and their Use in Science: Conceptions of Middle and High School Students and Experts. *Journal of Research in Science Teaching*, *28*(9), 799–822.

Grube, C. (2011). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung: Untersuchung der Struktur und Entwicklung Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I*. Dissertation. Zugriff am 01.10.2012. Verfügbar unter http://nbn-resolving.de/urn:nbn:de:hebis:34-2011041537247

Grünkorn, J. & Krüger, D. (2012). Entwicklung und Evaluierung von Aufgaben im offenen Antwortformat zur empirischen Überprüfung eines Kompetenzmodells zur Modellkompetenz. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (S. 9-27). Innsbruck: Studienverlag.

Grünkorn, J., Upmeier zu Belzen, A. & Krüger, D. (2011). Design and test of open-ended tasks to evaluate a theoretical structure of model competence. In A. Yarden & G. S. Carvalho (Hrsg.), *Authenticity in Biology Education. Benefits and Challenges. A selection of papers presented at the VIIIth Conference of European Researchers in Didactics of Biology (ERIDOB) Braga, Portugal*. (S. 53-65). Bragal: CIEC Universidade do Minho.

Grünkorn, J., A. Upmeier zu Belzen & D. Krüger (in Vorb.). *Students Perspectives on Models and Modelling Regarding a Theoretical Structure of Model Competence*.

Günther, J. (2006). *Lehrerfortbildung über die Natur der Naturwissenschaften: Studien über das Wissenschaftsverständnis von Grundschullehrkräften*. Berlin: Logos.

Günther, J., Grygier, P., Kircher, E., Sodian, B. & Thoermer, C. (2004). Studien zum Wissenschaftsverständnis von Grundschullehrkräften. In M. Prenzel (Hrsg.), *Bildungsqualität von Schule. Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (S. 93–113). Münster: Waxmann.

Haladyna, T. M. (1999). *Developing and Validating Multiple-Choice Test Items.* Mahwah, NJ: L. Erlbaum Associates.

Hammann, M. (2006). Kompetenzförderung und Aufgabenentwicklung. *MNU*, *59*(2), 85–95.

Hammann, M., Phan, T. T. H., Ehmer, M. & Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *MNU*, *59*(5), 292–299.

Harrison, A. G. & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education*, *22*(9), 1011–1026.

Hartig, J. (2008). Psychometric Models for the Assessment of Competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 69–90). Cambridge, MA: Hogrefe & Huber.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–171). Berlin: Springer.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Vorausetzungen technologiebasierter Kompetenzdiagnostik* (S. 17–36). Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 128–143). Berlin: Springer.

Hartmann, S., Upmeier zu Belzen, A. & Krüger, D. (2012, September). *KoWADiS. Modellierung und Erfassung naturwissenschaftlicher Kompetenzen bei Studierenden des Lehramts.* Poster auf der 77. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung, Bielefeld.

Held, C., Knauff, M. & Vosgerau, G. (2006). General Introduction: Current Developments in Cognitive Psychology, Neuroscience, and the Philosophy of Mind. In C. Held, M. Knauff & G. Vosgerau (Hrsg.), *Mental Models and the Mind. Current Developments in Cognitive Psychology, Neuroscience, and Philosophy of Mind* (S. 5–22). Amsterdam: Elsevier.

Heller, K. A. & Perleth, C. (2000). *KFT 4-12R - Kognitiver Fähigkeits-Test für 4. bis 12. Klassen, Revision*. Göttingen: Hogrefe.

Henze, I., Van Driel, J. H. & Verloop, N. (2007). Science teachers' knowledge about teaching models and modelling in the context of a new syllabus on public understanding of science. *Research in Science Education*, *37*, 99–122.

Hesse, M. B. (1966). *Models and analogies in science*. Notre Dame, IN: University of Notre Dame Press.

Hodson, D. (1993). Re-thinking Old Ways: Towards A More Critical Approach To Practical Work In School Science. *Studies in Science Education*, *22*, 85–142.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 28–72). Berlin: Springer.

Justi, R. S. & Gilbert, J. K. (2002). Modelling, teachers' views on the nature of modelling, and implications for the education of modellers. *International Journal of Science Education*, *24*(4), 369–387.

Justi, R. S. & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education*, *25*(11), 1369–1386.

Justi, R. S. & Gilbert, J. K. (2006). The role of analog models in the understanding of the nature of models in chemistry. In P. J. Aubusson, A. G. Harrison & S. M. Ritchie (Hrsg.), *Metaphor and Analogy in Science Education* (S. 119–130). Dordrecht: Springer.

Kattmann, U. (2006). Modelle. In H. Gropengießer (Hrsg.), Fachdidaktik Biologie: Die Biologiedidaktik (S. 330–339). Köln: Aulis.

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75–102). Berlin: Springer.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), Kompetenzdiagnostik. *Zeitschrift für Erziehungswissenschaft*, Sonderheft 8 (S. 11–29). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E., Hartig, J. & Rauch, D. (2008). The Concepts of Competence in Educational Contexts. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 3-22). Cambridge, MA: Hogrefe & Huber.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*(6), 876–903.

Klieme, E., Maag-Merki, K. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 5–15). Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2007). *Zur Entwicklung nationaler Bildungsstandards – Expertise*. Bonn Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E. & Steinert, B. (2004). Einführung der KMK-Bildungsstandards: Zielsetzungen, Konzeptionen und Einführung in den Schulen am Beispiel der Mathematik. *MNU*, *57*(3), 132–137.

Klos, S., Henke, C., Kieren, C., Walpuski, M. & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen - zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik*, *54*, 304–321.

KMK (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss: Beschluss vom 16.12.2004. Beschlüsse der Kultusministerkonferenz*. München: Luchterhand.

Knoblich, G. & Öllinger, M. (2006). Die Methode des Lauten Denkens. In J. Funke & P. A. Frensch (Hrsg.), *Handbuch der allgemeinen Psychologie - Kognition* (S. 691–696). Göttingen: Hogrefe.

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie / Journal of Psychology*, 216(2), 61–73.

Köller, O. (2008a). Bildungsstandards: Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik*, *54*(2), 163–173.

Köller, O. (2008b). Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaften*, Sonderheft 9, 47–59.

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. Cambridge, MA: MIT Press.

Kozma, R. B. & Russell, J. (1997). Multimedia and Understanding: Expert and Novice Responses to Different Representations of Chemical Phenomena. *Journal of Research in Science Teaching*, *34*(9), 949–968.

Kramer, G. (2009). *Entwicklung und Überprüfung eines Strukturmodells der fachlichen Kommunikationskompetenz im Biologieunterricht.* Dissertation. Zugriff am 01.10.2012. Verfügbar unter http://eldiss.uni-kiel.de/macau/servlets/MCRFileNodeServlet/dissertation_derivate_00003105/diss_gkramer.pdf?host=&o

Krell, M. & Krüger, D. (2010). Diagnose von Modellkompetenz: Deduktive Konstruktion und Selektion von geschlossenen Items. In D. Krüger, A. Upmeier zu Belzen & S. Nitz (Hrsg.), *Erkenntnisweg Biologiedidaktik 9* (S. 23–38). Kassel: Universität Kassel.

Krell, M. & Krüger, D. (2011). Forced Choice-Aufgaben zur Evaluation von Modellkompetenz im Biologieunterricht: Empirische Überprüfung konstrukt- und merkmalsbezogener Teilkompetenzen. In D. Krüger, A. Upmeier zu Belzen, P. Schmiemann & A. Sandmann (Hrsg.), *Erkenntnisweg Biologiedidaktik 10* (S. 53–68). Kassel: Univ. Kassel.

Krell, M. (2012). Using polytomous IRT models to evaluate theoretical levels of understanding models and modeling in biology education. *Science Education Review Letters*, 1-5.

Kremer, K., Grube, C., Urhahne, D. & Mayer, J. (2010). Exploring competencies in understanding the nature of science and scientific inquiry. In M. F. Taşar & G. Çakmakcı (Hrsg.), *Contemporary Science Education Research* (S. 245-254). Ankara, Turkey: Pegem Akademi.

Kubiszyn, T., & Borich, G. D. (2006). *Educational testing and measurement: Classroom application and practice.* Princeton, NJ: Wiley.

Lachmayer, S. (2008). *Entwicklung und Überprüfung eines Strukturmodells der Diagrammkompetenz für den Biologieunterricht.* Dissertation. Zugriff am

01.10.2012. Verfügbar unter http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_00003041

Lederman, N. G. (1992). Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research. *Journal of Research in Science Teaching*, *29*(4), 331–359.

Lederman, N. G. (2008). Nature of Science: Past, Present, and Future. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of Research on Science Education* (S. 831–879). New York: Routledge.

Lehrer, R. & Schauble, L. (2006). Scientific Thinking and Science Literacy. In K. A. Renninger, W. Damon, I. E. Sigel & R. M. Lerner (Hrsg.), *Handbook of child psychology* (S. 153–196). Hoboken, N.J: John Wiley & Sons.

Leisner, A. & Mikelskis, H. F. (2004). Erwerb metakonzeptueller Kompetenz durch ein systematisches Lernen über Modelle. In A. Pilon (Hrsg.), *Zur Didaktik der Physik und Chemie. Beitragsband zur Tagung in Berlin 2003* (S. 120–122). Münster: LiT.

Leisner-Bodenthin, A. (2006). Zur Entwicklung von Modellkompetenz im Physikunterricht. *ZfdN*, *12*, 91–108.

Leutner, D., Hartig, H. & Klieme, E. (2008). Measuring Competencies: Introduction to Concepts and Questions of Assessment in Education. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 190–205). Cambridge, MA: Hogrefe & Huber.

Leutner, D., Klieme, E., Meyer, K. & Wirth, J. (2004). Problemlösen. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des Zweiten Internationalen Vergleichs* (S. 147–175). Weinheim: Beltz.

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2007). Landesweite Lernstandserhebungen Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. *Zeitschrift für Erziehungswissenschaften*, Sonderheft 8, 149–167.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Lind, G., Friege, G., Kleinschmidt, L. & Sandmann, A. (2004). Beispiellernen und Problemlösen. *ZfdN*, *10*, 29–49.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Lüdtke, O. & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke & L. Stecher (Hrsg.), *Enzyklopädie Erziehungswissenschaft Online* (S. 1–42). Weinheim: Juventa.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, *58*(2), 103–117.

Mahr, B. (2003). Modellieren: Beobachtungen und Gedanken zur Geschichte des Modellbegriffs. In S. Krämer & H. Bredekamp (Hrsg.), *Bild, Schrift, Zahl* (S. 59–86). München: Fink.

Mahr, B. (2004). *Das Wissen im Modell* (KIT-Report Nr. 150). TU-Berlin. Zugriff am 01.10.2012. Verfügbar unter http://www.flp.tu-berlin.de/fileadmin/fg53/KIT-Reports/r150.pdf

Mahr, B. (2008a). Ein Modell des Modellseins: Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks & E. Knobloch (Hrsg.), *Modelle* (S. 187–218). Frankfurt a. M.: Peter Lang.

Mahr, B. (2008b). Cargo: Zum Verhältnis von Bild und Modell. In I. Reichle (Hrsg.), *Visuelle Modelle* (S. 17–40). München: Fink.

Mahr, B. (2009). Die Informatik und die Logik der Modelle. *Informatik Spektrum*, 228–249.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 177–186). Berlin: Springer.

Mayring, P. (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim: Beltz.

McClelland, D. C. (1973). Testing for Competence Rather Than for "Intelligence". *American Psychologist*, *28*, 1–14.

Meisert, A. (2008). Vom Modellwissen zum Modellverständnis: Elemente einer umfassenden Modellkompetenz und deren Fundierung durch lernerseitige Kriterien zur Klassifikation von Modellen. *ZfdN*, *14*, 243–261.

Meisert, A. (2009). Modelle in der Biologie: Wie lässt sich im Unterricht ein Verständnis für ihre Bedeutung fördern? *MNU*, *62*(7), 424–430.

Mislevy, R. J. & Haertel, G. D. (2006). *PADI Technical Report 17: Implications of Evidence-Centered Design for Educational Testing*. Menlo Park, CA: SRI International.

Mittelstraß, J. (2004). *Enzyklopädie Philosophie und Wissenschaftstheorie*. Stuttgart: Metzler.

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227–274). Berlin: Springer.

Morrison, M. & Morgan, M. S. (1999a). Models as mediating instruments. In M. S. Morgan & M. Morrison (Hrsg.), *Models as mediators. Perspectives on natural and social sciences* (S. 10–37). Cambridge, NY: Cambridge University Press.

Morrison, M. & Morgan, M. S. (1999b). Introduction. In M. S. Morgan & M. Morrison (Hrsg.), *Models as mediators. Perspectives on natural and social sciences* (S. 1–9). Cambridge, NY: Cambridge University Press.

Muthén, L. K. & Muthén, B. O. (2007). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

National Research Council (Hrsg.). (1996). *National Science Education Standards.* Washington, DC: The National Academies Press.

Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Hrsg.), *Cognitive models of science* (S. 3–45). Minneapolis, MN: University of Minnesota Press.

Nersessian, N. J. (1999). Model-Based Reasoning in Conceptual Change. In L. Magnani, N. J. Nersessian & P. Thagard (Hrsg.), *Model-based reasoning in scientific discovery* (S. 5–22). New York: Kluwer Academic/Plenum Publishers.

Neuhaus, B. & Braun, E. (2007). Testkonstruktion und Testanalyse – praktische Tipps für empirisch arbeitende Didaktiker und Schulpraktiker. In H. Bayrhuber, D. Elster, D. Krüger & H. J. Vollmer (Hrsg.), *Forschungen zur Fachdidaktik: Kompetenzentwicklung und Assessment* (S. 135–164). Innsbruck: StudienVerlag.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Hrsg.), *Mental models* (S. 7–14). Hillsdale, NJ: Lawrence Erlbaum.

Nowak, K., Nehring, A., Tiemann, R. & Upmeier zu Belzen, A. (2012, September). *Focus on Processes of Scientific Inquiry in Theory and Practice*. Poster auf der 9. Conference of European Researchers in Didactics of Biology (Eridob), Berlin.

Oh, P. S. & Oh, S. J. (2011). What Teachers of Science Need to Know about Models: An overview. *International Journal of Science Education*, *33*(8), 1109–1130.

Orsenne, J. & Upmeier zu Belzen, A. (2012). Hands-On Aufgaben zur Erfassung und Förderung von Modellkompetenz im Biologieunterricht. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (S. 29–44). Innsbruck: Studienverlag.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R. & Duschl, R. (2003). What "Ideas-about-Science" Should Be Taught in School Science?: A Delphi Study of the Expert Community. *Journal of Research in Science Teaching*, *40*(7), 692–720.

Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* Boston: Kluwer Academic Publishers.

Patzke, C. (2010). *Validierung von Aufgaben zur Diagnose kognitiver Prozesse bei Schülerinnen und Schülern der 10. Jahrgangsstufe*. Unveröffentlichte Wissenschaftliche Hausarbeit zur Ersten Staatsprüfung für das Amt des Studienrats, Humboldt-Universität zu Berlin.

Patzke, C. & Upmeier zu Belzen, A. (2011). Entwicklung von Modellkompetenz im Biologieunterricht [Abstract]. In F. X. Bogner (Hrsg.), *Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBIOo*. (S. 168–169). Bayreuth: Difo Druck.

Patzke, C. & Upmeier zu Belzen, A. (2012). Entwicklung von Modellkompetenz und entsprechende Lerngelegenheiten im Biologieunterricht [Abstract]. In D. Elster, A. Schultz-Siatkowski & F. Wischmann (Hrsg.), *Tagungsband der 14.*

Frühjahrsschule der Fachsektion Didaktik in Biologie im VBIO (S. 104–105). Aachen: Shaker-Verlag.

Peugh, J. L. & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, *74*(4), 525–556.

Prenzel, M., Carstensen, C. H., Frey, A., Drechsel, B. & Rönnebeck, S. (2007). PISA 2006: Eine Einführung in die Studie. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 31–59). Münster: Waxmann.

Prenzel, M., Drechsel, B., Carstensen, C. H. & Ramm, G. (2004). PISA 2003 – eine Einführung. In M. Prenzel, J. Baumert, W. Blum, R. H. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H., G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 13-46). Münster: Waxmann.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, *30*(1), 120–135.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63–106). Münster: Waxmann.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63–106). Münster: Waxmann.

Rauch, D. & Hartig, J. (2012). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 253–263). Berlin: Springer.

Reckase, M. D. (1997). The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement*, *21*(1), 25–36.

Reitschert, K. & Hößle, C. (2007). Wie Schüler ethisch bewerten: Eine qualitative Untersuchung zur Strukturierung und Ausdifferenzierung von Bewertungskompetenz in bioethischen Sachverhalten bei Schülern der Sek I. *ZfdN*, *13*, 125–143.

Rösch, H. (2003). *Deutsch als Zweitsprache: Grundlagen, Übungsideen, Kopiervorlagen zur Sprachförderung*. Hannover: Schroedel.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Rost, J., Prenzel, M., Carstensen, C. H., Senkbeil, M. & Groß, K. (2004). *Naturwissenschaftliche Bildung in Deutschland: Methoden und Ergebnisse von PISA 2000.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *ZfdN*, *12*, 45–66.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–142). Berlin, Heidelberg: Springer.

Schmiemann, P. (2010). *Modellierung von Schülerkompetenzen im Bereich des biologischen Fachwissens.* Berlin: Logos.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12: Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12*. Göttingen: Hogrefe.

Schütte, K., Frenzel, A. C., Asseburg, R. & Pekrun, R. (2007). Schülermerkmale, naturwissenschaftliche Kompetenz und Berufserwartung. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 125–146). Münster: Waxmann.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., et al. (2009). Developing a Learning Progression for Scientific Modelling: Making Scientific Modelling Accessible and Meaningful for Learners. *Journal of Research in Science Teaching*, *46*(6), 632–654.

Schwarz, C. V. & Gwekwerere, Y. N. (2007). Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education*, *91*(1), 158–186.

Schwarz, C. V. & White, B. Y. (2005). Metamodeling Knowledge: Developing Students' Understanding of Scientific Modeling. *Cognition and Instruction*, *23*(2), 165–203.

Sedlmeier, P. & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson Studium.

Senatsverwaltung für Bildung, Jugend und Sport (2006). *Rahmenlehrplan für die Sekundarstufe I: Biologie*. Berlin: Oktoberdruck AG.

Senkbeil, M., Rost, J., Carstensen, C. H., & Walter, O. (2005). Der nationale Naturwissenschaftstest PISA 2003: Entwicklung und empirische Überprüfung eines zweidimensionalen Facettendesigns. *Empirische Pädagogik*, *19*(2), 166–189.

Sins, P. H. M., Savelsbergh, E. R., Joolingen, W. R. van & Hout-Wolters, B. H. A. M. van. (2009). The Relation between Students' Epistemological Understanding of Computer Models and their Cognitive Processing on a Modelling Task. *International Journal of Science Education*, *31*(9), 1205–1229.

Sommer, C. (2006). *Untersuchung der Systemkompetenz von Grundschülern im Bereich Biologie*. Dissertation. Zugriff am 01.10.2012. Verfügbar unter http://eldiss.uni-kiel.de/macau/servlets/MCRFileNodeServlet/dissertation_derivate_00001652/d1652.pdf?host=&o

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien: Springer.

Stachowiak, H. (1983). Erkenntnisstufen zum systematischen Neopragmatismus und zur Allgemeinen Modelltheorie. In H. Stachowiak (Hrsg.), *Modelle. Konstruktion der Wirklichkeit* (S. 87–146). München: Fink.

Stephens, S.-A., McRobbie, C. J. & Lucas, K. B. (1999). Model-Based Reasoning in a Year 10 Classroom. *Research in Science Education*, *29*(2), 189–208.

Suárez, M. (1999). Theories, Models, and Representations. In L. Magnani, N. J. Nersessian & P. Thagard (Hrsg.), *Model-based reasoning in scientific discovery* (S. 75–84). New York: Kluwer Academic/Plenum Publishers.

Suppes, P. (1961). A comparison of the meaning and uses of models in mathematics and the empirical sciences. In H. von Freudenthal (Hrsg.), *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences* (S. 163–177). Dordrecht: Reidel.

Terzer, E., Hartig, J. & Upmeier zu Belzen, A. (mit Änd. angen.). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *ZfdN*.

Terzer, E., Patzke, C. & Upmeier zu Belzen, A. (2012). Validierung von Multiple-Choice Items zur Modellkompetenz durch lautes Denken. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (S. 45–62). Innsbruck: Studienverlag.

Trautwein, U., Lüdtke, O., Becker, M., Neumann, M. & Nagy, G. (2008). Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. In E. Schlemmer & H. Gerstberger (Hrsg.), *Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis* (S. 91–107). Wiesbaden: VS Verlag für Sozialwissenschaft.

Treagust, D. F., Chittleborough, G. & Mamiala, T. L. (2001). Students' concept of models: An epistemological and ontological perspective. *Proceedings Western Australian Institute for Educational Research Forum 2000*. Zugriff am 01.10.2012. Verfügbar unter http://www.waier.org.au/forums/2001/treagust.html

Treagust, D. F., Chittleborough, G. D. & Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. *International Journal of Science Education*, *24*(4), 357–368.

Treagust, D. F., Chittleborough, G. D. & Mamiala, T. L. (2004). Students' Understanding of the Descriptive and Predictive Nature of Teaching Models in Organic Chemistry. *Research in Science Education*, *34*, 1–20.

Trier, U. & Upmeier zu Belzen, A. (2009). „Die Wissenschaftler nutzen Modelle, um etwas Neues zu entdecken, und in der Schule lernt man einfach nur, dass es so ist.": Schülervorstellungen zu Modellen. In D. Krüger, A. Upmeier zu Belzen, S. Hof, K. Kremer & J. Mayer (Hrsg.), *Erkenntnisweg Biologiedidaktik 8* (S. 23–37). Kassel: Universität Kassel.

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *ZfdN*, *16*, 41–57.

Urhahne, D., Kremer, K. & Mayer, J. (2008). Welches Verständnis haben Jugendliche von der Natur der Naturwissenschaften? Entwicklung und erste Schritte zur Validierung eines Fragebogens. *Unterrichtswissenschaft*, *36*, 72–94.

Urhahne, D., Kremer, K., Mayer, J. & Mayer, J. (2011). Conceptions of the nature of science: Are they general or context-specific? *International Journal of Science and Mathematics Education*, *9*(3), 707–730.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

van Buuren, S. & Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: Netherlands Organization for Applied Scientific (TNO).

Van Driel, J. H. & Verloop, N. (1999). Teachers' knowledge of models and modelling in Science. *International Journal of Science Education*, *21*(11), 1141–1153.

van Fraassen, B. C. (1980). *The scientific image*. New York: Clarendon Press.

Veenman, M. V. J. (2005). The Assessment of Metacognitive Skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Hrsg.), *Lernstrategien und Metakognition. Implikationen für Forschung und Praxis* (S. 77–99). Münster: Waxmann.

Volodin, N. A. & Adams, R. J. (1995). *Identifying and estimating a D-dimensional item response model.* Paper presented at the International Objective Measurement Workshop. Berkeley, CA: University of California.

Vonken, M. (2005). *Handlung und Kompetenz: Theoretische Perspektiven für die Erwachsenen- und Berufspädagogik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Vosgerau, G. (2006). The Perceptual Nature of Mental Models. In C. Held, M. Knauff & G. Vosgerau (Hrsg.), *Mental Models and the Mind. Current Developments in Cognitive Psychology, Neuroscience, and Philosophy of Mind* (S. 255–275). Amsterdam: Elsevier.

Vosniadou, S. & Ioannides, C. (1998). From conceptual development to science education: a psychological point of view. *International Journal of Science Education*, *20*(10), 1213–1230.

Vosniadou, S. (2002). Mental Models in Conceptual Development. In L. Magnani & N. J. Nersessian (Hrsg.), *Model-based reasoning. Science, technology, values* (S. 353–368). New York: Kluwer Academic.

Weidle, R. & Wagner, A. C. (1994). Die Methode des Lauten Denkens. In G. L. Huber & H. Mandl (Hrsg.), *Verbale Daten. Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung* (S. 81–103). Weinheim: Beltz.

Weinert, F. E. (2001). Concepts of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Seattle, WA: Hogrefe & Huber.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung*. Berlin: Logos.

White, R. W. (1959). Motivation Reconsidered: The Concept of Competence. *Psychological Review*, *66*(4), 297–331.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wirtz, M. A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wu, M. (2005). The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Wu, M. L., Adams, R. J. & Wilson, M. R. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software*. Camberwell, VIC: ACER Press.

Zöfel, P. (2003). *Statistik für Psychologen*. München: Pearson Studium.

Anhang

Anhang 1 – Kategorisierung von Vorstellungen zu Modellen in empirischen Untersuchungen.

Studie	**Kategorien**				
AAAS	Geometric figures, number sequences, graphs, diagrams, sketches, number lines, maps, and oral and written descriptions can be used to represent objects, events, and processes in the real world.	A model of something is similar to but not exactly like the thing being modeled. There is no guarantee that ideas based solely on a model are correct.		Models are useful for thinking about real-world objects, events, and processes. The usefulness of a model in thinking about objects, events, and processes depends on how closely its behavior matches key aspects of what is being modeled.	

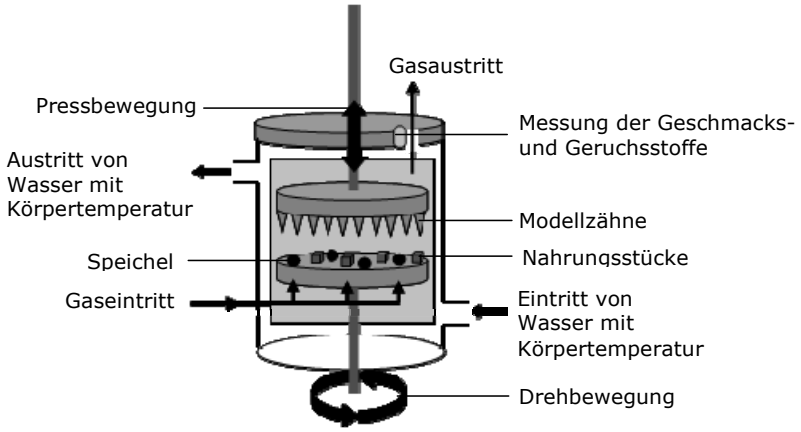
Studie	**Kategorien**							
Crawford & Cullin (2005)			multiple models for the same thing	purpose of models	validating/testing models	designing and creating models	changing a model	
Grosslight et al. (1991)	kinds of models		multiple models	purpose of models		designing and creating models	changing a model	
Justi & Gilbert (2003)	entities of which a model consists	nature of a model	its relative uniqueness	its status in the making of a prediction	use to which it can be put		time span over which it is used	basis of accreditation for its existence
Meisert (2009)		Modell-Original-Relation		Funktion von Modellen als Mittel der Erkenntnisgewinnung		Entwicklungs-charakter von Modellen	Rolle des Modellierers	
Schwarz & White (2005)	nature of models			purpose or utility of models	evaluation of models	nature or process of modeling		
Schwarz et al. (2009)			scientific models as tools for predicting and explaning					
		models change as understanding improves						

Studie	**Kategorien**									
Sins et al. (2009)	nature of models		purposes of models	evaluation of models		process of modelling				Sins et al. (2009)
Meisert (2009)		Modell-Original-Relation		Funktion von Modellen als Mittel der Erkenntnisgewinnung			Entwicklungscharakter von Modellen	Rolle des Modellierers		
Treagust et al. (2001) - VOMS	representations of ideas or how things work; accurate duplicates of reality		one model only; many models			facts that support the model and the theory; personal feelings or motives; can be used successfully to explain results			change	support by a large majority of scientists
Treagust et al. (2002) - SUMS	models as exact replicas		scientific models as multiple representations	models as explanatory tools	how scientific models are used				changing nature of scientific models	
Van Driel & Verloop (1999)	types of representations of models	characteristics of scientific models		goals and functions of models in science		modelling in science (design and revision)				

Anhang 2 – Itembeschreibungen in der Konstruktionsanleitung.

Beispiel für einen Aufgabenstamm – Der (K)Automat

Französische Wissenschaftler haben einen künstlichen Mund entwickelt. Mit diesem Modell kann die Verteilung von Geschmacks- und Geruchsstoffen bei einem kauenden Menschen nachgeahmt werden. Das Modell kann typische Kaubewegungen ausführen und arbeitet mit künstlichem Speichelfluss bei Körpertemperatur. Mithilfe eines Gases, welches durch das Modell strömt, werden frei werdende Geschmacks- und Geruchsstoffe aus dem Apparat nach außen transportiert, wo sie dann genau untersucht werden können. Die Wissenschaftler können aus diesen Untersuchungen ableiten, was genau beim Zerkauen der Nahrung unter Einwirkung von Speichel im Mund passiert. Der künstliche Mund könnte zukünftig bei der elektronischen Analyse von Lebensmitteln helfen.



(K)Automat.

Eigenschaften von Modellen, Niveau I - Modelle sind Kopien *von* etwas

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**In welcher Eigenschaft stimmt [das Modell] mit [dem Original] überein?** Beispiel: In welcher Eigenschaft stimmt der (K)Automat mit der Vorverdauung im Mund überein?
Antwortmöglichkeiten	**Aspekte des Modellobjekts** (Farbe, Form, Größe, Material etc.)
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, dass der (K)Automat einen Mund in Bezug auf Zähne, Kaubewegung, Körpertemperatur, Speichelfluss, Nahrung, Geschmacks- und Geruchsstoffe nachahmt.

Eigenschaften von Modellen, Niveau II - Modelle sind idealisierte Repräsentationen *von* etwas

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Welche Eigenschaften [des Originals] sind [im Modell] vereinfacht?** Beispiel: Welche Eigenschaften der Vorverdauung im Mund sind im (K)Automaten vereinfacht?
Antwortmöglichkeiten	**Eigenschaften des Originals**
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, dass der (K)Automat den Mund idealisiert darstellt und insbesondere die Struktureigenschaften (Farbe, exakte Form, einzelne Bestandteile) dabei vernachlässigt werden.

Eigenschaften von Modellen, Niveau III - Modelle sind theoretische Rekonstruktionen *von* etwas

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Welche Annahme über [das Original] war die Grundlage für die Entwicklung [des Modells]?** Beispiel: Welche Annahme über die Vorverdauung im Mund war die Grundlage für die Entwicklung des (K)Automaten?
Antwortmöglichkeiten	denkbare **Annahmen**
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, welche Vorstellungen über die Vorverdauung im Mund der Konstruktion des (K)Automaten zugrunde liegen. Somit erkennen sie Modelle als theoriegeleitet.

Alternative Modelle, Niveau I - Unterschiede zwischen den *Modellobjekten*

Stamm	**Darstellung von Modellen**, die sich durch **verschiedene Objekteigenschaften** (Farbe, Größe, Perspektive, Material etc.) unterscheiden, z. B. verschiedene (K)Automat-Versionen **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Warum gibt es unterschiedliche Modelle zu [dem Original]?** Beispiel: Warum gibt es unterschiedliche Modelle zur Vorverdauung im Mund?
Antwort-möglichkeiten	denkbare Gründe – Argumentation bezieht sich nur auf **Modellobjekt**
gemessene Kompetenz	Die Schülerinnen und Schüler begründen die Existenz der (K)Automatversionen mit unterschiedlichen Materialien, Größen, Dimensionalität (2D, 3D) etc.

Alternative Modelle, Niveau II - Ausgangsobjekt ermöglicht Herstellung verschiedener Modelle *von* etwas

Stamm	**Darstellung von Modellen**, die sich durch den Fokus auf **verschiedene Aspekte des Originals** unterscheiden, z. B. (K)Automat und Computersimulation zum Zusammenspiel der Muskeln bei der Kaubewegung **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Warum gibt es unterschiedliche Modelle zu [dem Original]?** Beispiel: Warum gibt es unterschiedliche Modelle zur Vorverdauung im Mund?
Antwort-möglichkeiten	denkbare Gründe – Argumentation bezieht sich auf den **Fokus auf das Original**
gemessene Kompetenz	Die Schülerinnen und Schüler begründen die Existenz der (K)Automatversionen mit dem Fokus, der den Versionen zugrunde liegt (Kaubewegung als muskulärer Prozess bzw. als Teil der Vorverdauung).

Alternative Modelle, Niveau III - Modelle für verschie-dene Hypothesen

Stamm	**Darstellung von Modellen**, die sich durch **verschiedene Hypothesen**, die ihnen zugrundeliegen, unterscheiden, z. B. (K)Automat mit und ohne Zunge (Hypothese 1: Die Zunge spielt eine Rolle bei der Vorverdauung von Nahrung; Hypothese 2: Die Zunge spielt keine Rolle bei der Vorverdauung) **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Warum gibt es unterschiedliche Modelle zu [dem Original]?** Beispiel: Warum gibt es unterschiedliche Modelle zur Vorverdauung im Mund?
Antwort-möglichkeiten	denkbare Gründe – Argumentation bezieht sich auf **Forschungsprozess**
gemessene Kompetenz	Die Schülerinnen und Schüler begründen die Existenz der (K)Automatversionen mit den Hypothesen, die den Versionen zugrunde liegen.

Zweck von Modellen, Niveau I - Modellobjekt zur Beschreibung *von* etwas einsetzen

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Was kann [man] mit [dem Modell] zeigen?** Beispiel: Was kann man mit dem (K)Automaten zeigen?
Antwort-möglichkeiten	denkbare **Modellinhalte** (hier keine Zusammenhänge)
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, dass der Kautomat den Mund als Vorverdauungsorgan veranschaulicht.

Zweck von Modellen, Niveau II - Bekannte Zusammenhänge und Korrelationen *von* Variablen im Ausgangsobjekt erklären

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Welchen Zusammenhang kann man mit [dem Modell] erklären?** Beispiel: Welchen Zusammenhang kann man mit dem (K)Automaten erklären?
Antwort-möglichkeiten	denkbare **bekannte Zusammenhänge im Original**
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, dass der (K)Automat korrelative Zusammenhänge zwischen Kaubewegung, Speichelfluss und Körpertemperatur erklärt.

Zweck von Modellen, Niveau III - Zusammenhänge von Variablen *für* zukünftige neue Erkenntnisse voraussagen

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Daten aus Modellexperiment**, z. B. zu Kaubewegung und Speichelfluss
Frage	**Welche Vermutung über [das Original] kann [man] aus [dem Modell] ableiten?** Beispiel: Welche Vermutung über die Vorverdauung im Mund können die Wissenschaftler von dem (K)Automaten ableiten?
Antwort-möglichkeiten	denkbare **Hypothesen**
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, inwiefern der (K)Automat Hypothesen über Kausalzusammenhänge zwischen Kaubewegung, Speichelfluss, Körpertemperatur und der Verteilung von Geschmacks- und Geruchsstoffen zulässt.

Testen von Modellen, Niveau I - *Modellobjekt* überprüfen

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund
Frage	**Wie kann [man] prüfen, ob [man] [das Modell] einsetzen kann?** Beispiel: Wie können die Wissenschaftler prüfen, ob sie den (K)Automaten einsetzen können?
Antwort-möglichkeiten	denkbare **Prüfverfahren zur Einsetzbarkeit des Modellobjekts**
gemessene Kompetenz	Die Schülerinnen und Schüler benennen, wie die Einsetzbarkeit des (K)Automaten geprüft werden kann, z. B. Funktionalität der Kaubewegung, Speichelfluss, Gasfluss, Einstellung der Temperatur etc.

Testen von Modellen, Niveau II - Parallelisieren mit dem Ausgangsobjekt, Modell *von* etwas testen

Stamm	**Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Zweck eines entsprechenden Modells**, z. B. des (K)Automats
Frage	**Welches dieser Modelle kann für [den Zweck] genutzt werden?** Beispiel: Welches dieser Modelle kann zur Veranschaulichung der Vorverdauung genutzt werden?
Antwort-möglichkeiten	denkbare **Modelle**, die sich in der Angemessenheit der Parallelen zwischen Modell und Original (in Bezug auf Struktur und/oder Funktion) für den Zweck unterscheiden
gemessene Kompetenz	Die Schülerinnen und Schüler benennen ein Modell, das eine adäqaute Passung zum Mund aufweist, und prüfen insofern die Tauglichkeit des Modells.

Testen von Modellen, Niveau III - Überprüfen von Hypothesen bei der Anwendung, Modell *für* etwas testen

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Hypothese** (= Grundlage für Modell), z. B. Die mechanische Zerkleinerung von Nahrung, der Speichel sowie die Körpertemperatur beeinflussen die Verteilung der Geschmacks- und Geruchsstoffe im Mund.
Frage	**Wie kann man [diese Vermutung] testen?** Beispiel: Wie können die Wissenschaftler diese Vermutung testen?
Antwort-möglichkeiten	denkbare **Versuchsdesigns**
gemessene Kompetenz	Die Schülerinnen und Schüler erkennen, welches experimentelle Design zur Prüfung der Hypothese, dass die mechanische Zerkleinerung von Nahrung, der Speichel sowie die Körpertemperatur die Verteilung der Geschmacks- und Geruchsstoffe im Mund beeinflussen, mit dem (K)Automat geeignet ist.

Ändern von Modellen, Niveau I - Mängel am *Modellobjekt* beheben

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Zweck des Modells**, z. B. Analyse von Geschmacks- und Geruchsstoffen **Fehler im Modellobjekt**, z. B. Speichelfluss funktioniert nicht
Frage	**Was muss [man] an [diesem Modell] verändern, damit [der Zweck erfüllt ist]?** Beispiel: Welche Veränderungen müssen die Wissenschaftler zur Analyse von Geschmacks- und Geruchsstoffen am (K)Automaten vornehmen?
Antwort-möglichkeiten	denkbare **Veränderungen des Modellobjekts**
gemessene Kompetenz	Die Schülerinnen und Schüler benennen, *wie* z. B. der fehlerhafte Speichelfluss am (K)Automaten behoben werden kann.

Ändern von Modellen, Niveau II - Modell als Modell *von* etwas durch neue Erkenntnisse oder zusätzliche Perspektiven revidieren

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Zweck des Modells**, z. B. Analyse von Geschmacks- und Geruchsstoffen **neue Erkenntnisse**, z. B. Relevanz eines weiteren Vorverdauungsfaktors im Mund wie etwa neu entdeckte Bakterien
Frage	**Was muss [man] nach [diesen Erkenntnissen] an [diesem Modell] verändern?** Beispiel: Was müssen die Wissenschaftler nach der Entdeckung unbekannter Bakterien am (K)Automaten verändern?
Antwort-möglichkeiten	denkbare **Veränderungen**
gemessene Kompetenz	Die Schülerinnen und Schüler benennen, welche Veränderungen wegen dieser Entdeckung am (K)Automaten vorgenommen werden müssen, damit der (K)Automat seinen Zweck noch erfüllt.

Ändern von Modellen, Niveau III - Modell *für* etwas aufgrund falsifizierter Hypothesen revidieren

Stamm	**Darstellung des Modells**, z. B. des (K)Automats **Fachwissen zum Original**, z. B. zur Vorverdauung im Mund **Zweck des Modells**, z. B. Analyse von Geschmacks- und Geruchsstoffen **Daten aus Modellexperiment**, z. B. zur Übereinstimmung von Kaubewegung von (K)Automat und Mund
Frage	**Die Vermutung [im Modellexperiment] stimmt nicht mit [dem Original] überein. Was muss [man] deshalb an [diesem Modell] verändern?** Beispiel: Die Ergebnisse aus diesem Experiment passen nicht zum (K)Automaten. Welche Veränderungen müssen die Wissenschaftler auf der Grundlage dieser Daten am (K)Automaten vornehmen?
Antwort-möglichkeiten	denkbare **Veränderungen**
gemessene Kompetenz	Die Schülerinnen und Schüler benennen, welche Veränderungen wegen dieser Daten am (K)Automaten vorgenommen werden müssen, damit der (K)Automat seinen Zweck noch erfüllt.

Anhang 3 – Rohdaten zum Rating der Konstruktionsanleitung.

Rohdaten des Ratings der Konstruktionsanleitung. R = Raterin bzw. Rater, E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die Nummerierung 1 bis 3 bei den Zuordnungen kennzeichnet die Niveaustufe. Von der theoretischen abweichende Zuordnungen sind grau markiert.

	Theorie	R1	R2	R3	R4	R5	R6	R7	R8	R9
Itembeschreibung 1	*Ä2*	Ä2	Ä2	Ä2		Ä2		Ä2	**Ä3**	
Itembeschreibung 2	*E1*	**E2**	E1	E1		E1		E1	**T2**	
Itembeschreibung 3	*T3*	T3	T3	T3		T3		T3	T3	
Itembeschreibung 4	*E3*	E3	E3	E3		E3		**E2**	E3	
Itembeschreibung 5	*A2*	A2	**A1**	**Z2**		A2		A2	A2	
Itembeschreibung 6	*Z1*		**E1**		Z1		Z1	**E1**	Z1	Z1
Itembeschreibung 7	*A3*		A3		A3		A3	A3	A3	A3
Itembeschreibung 8	*T2*		**E2**		T2		**T1**	T2	T2	T2
Itembeschreibung 9	*Ä1*		Ä1		Ä1		Ä1	Ä1	Ä1	Ä1
Itembeschreibung 10	*Z3*		Z3		Z3		Z3	Z3	Z3	Z3
Itembeschreibung 11	*T1*	T1		T1	**T2**	T1	**T2**			T1
Itembeschreibung 12	*Z2*	Z2		Z2	Z2	Z2	Z2			Z2
Itembeschreibung 13	*Ä3*	Ä3		**Ä2**	**Ä2**	Ä3	Ä3			Ä3
Itembeschreibung 14	*A1*	A1		A1	A1	A1	A1			A1
Itembeschreibung 15	*E2*	**E1**		E2	**E1**	E2	E2			E2

Anhang 4 – Überblick über die psychometrische Qualität der Items in der Itemerprobung.

Überblick über die psychometrische Qualität der Items zur Dimension ‚Kenntnisse über Modelle'. Da die Items in Teilstudien geprüft wurden und somit eine Skalierung je Teilkompetenz nicht möglich war, ist hier die Trennschärfe in Bezug auf Modellkompetenz angegeben. Diese kann nur als grobe Orientierung dienen. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die zweite Stelle der Itembezeichnung (Nummerierung 1 bis 3) kennzeichnet die Niveaustufe, die dritte die Itemnummer.

Item	n	Schwierigkeit	Distraktorenanalyse			wMNSQ	T	Trennschärfe
E1.3	92	-0.598	12	61	27	0.98	-0.3	.51
E1.4	95	-2.535	22	22	56	0.90	-0.3	.56
E1.6	58	0.034	46	18	36	1.01	0.1	.38
E2.1	89	-0.396	22	31	47	0.94	-0.8	.65
E2.3	56	1.762	30	32	39	0.98	0.0	.39
E2.4	56	0.306	48	32	19	1.08	1.0	.28
E3.3	89	0.044	20	49	31	1.07	0.9	.42
E3.4	90	-0.204	61	17	22	0.91	-1.4	.64
E3.6	59	0.082	21	11	68	1.01	0.2	.41
A1.1	101	-0.088	13	32	55	1.01	0.1	.42
A1.2	56	-0.725	22	17	61	0.92	-0.7	.55
A1.3	58	-0.490	24	29	48	0.92	-0.9	.59
A2.1	85	0.366	34	28	38	0.96	-0.5	.54
A2.2	88	-0.907	27	38	35	0.99	-0.1	.40
A2.3	87	-0.147	15	46	38	0.99	-0.1	.56
A3.1	56	0.465	19	44	38	0.89	-1.2	.60
A3.2	58	0.814	16	30	54	1.05	0.4	.28
A3.5	59	0.448	29	26	44	1.07	0.9	.24

Überblick über die psychometrische Qualität der Items zur Dimension ‚Modellbildung'. Da die Items in Teilstudien geprüft wurden und somit eine Skalierung je Teilkompetenz nicht möglich war, ist hier die Trennschärfe in Bezug auf Modellkompetenz angegeben. Diese kann nur als grobe Orientierung dienen. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die zweite Stelle der

Itembezeichnung (Nummerierung 1 bis 3) kennzeichnet die Niveaustufe, die dritte die Itemnummer.

Item	n	Schwierigkeit	Distraktorenanalyse			wMNSQ	T	Trennschärfe
Z1.3	89	-1.071	57	35	9	1.04	0.3	.35
Z1.4	90	-1.011	36	8	56	1.03	0.3	.50
Z1.1	58	-0.894	39	44	17	0.96	-0.4	.49
Z2.5	82	-0.673	27	19	54	0.98	-0.2	.47
Z2.6	84	0.381	40	52	8	1.09	1.1	.38
Z2.7	85	-0.401	35	9	56	1.05	0.7	.42
Z3.1	85	-0.024	29	7	64	1.02	0.3	.44
Z3.3	86	-0.448	20	37	43	0.92	-0.9	.61
Z3.4	83	1.105	22	13	65	1.09	0.7	.29
T1.1	102	-0.583	41	30	30	0.98	-0.2	.54
T1.2	100	0.220	51	13	36	1.15	1.1	.34
T1.4	59	-0.136	43	22	35	0.98	-0.3	.55
T2.3	73	-1.380	56	25	19	1.00	0.0	.39
T2.5	58	-0.275	33	33	33	0.99	-0.2	.38
T2.6	58	-0.894	41	24	35	0.97	-0.3	.46
T3.1	89	-0.955	21	50	29	0.94	-0.5	.58
T3.3	90	0.341	18	64	18	0.97	-0.4	.50
T3.4	89	-0.306	58	28	14	1.02	0.3	.55
Ä1.3	90	-0.766	30	53	13	1.02	0.2	.42
Ä1.5	91	-1.513	53	32	16	0.96	-0.3	.50
Ä1.6	90	-0.245	43	40	18	0.94	-0.9	.60
Ä2.1	88	0.340	40	38	22	0.92	-1.1	.61
Ä2.3	87	-0.427	12	47	41	0.97	-0.3	.57
Ä2.4	89	-0.862	32	21	46	1.02	0.2	.40
Ä3.1	91	0.880	21	22	57	1.18	1.7	.13
Ä3.4	56	0.385	17	43	40	1.01	0.1	.43
Ä3.5	59	0.282	35	45	19	1.01	0.1	.33

Anhang 5 – Rohdaten zum Rating der einzelnen Items.

Rohdaten des Ratings der Items 1-20. R = Raterin bzw. Rater, E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die Nummerierung 1 bis 3 bei den Zuordnungen kennzeichnet die Niveaustufe. Abweichungen von der theoretischen Zuordnung sind grau markiert.

	Theorie	R1	R2	R3	R4	R9
Item 1: A1.1	*A1*	**A2**				**A2**
Item 2: E3.6	*E3*	**T3**				**Z2**
Item 3: Ä2.4	*Ä2*	Ä2				Ä2
Item 4: Z1.2	*Z1*	Z1				Z1
Item 5: E1.6	*E1*	**E2**				**E2**
Item 6: A3.5	*A3*	A3	**Z2**			
Item 7: A2.1	*A2*	**A1**	A2			
Item 8: T3.1	*T3*	T3	T3			
Item 9: T2.6	*T2*	T2	**Z1**			
Item 10: Z3.3	*Z3*	**A3**	**T3**			
Item 11: Ä3.1	*Ä3*		Ä3	Ä3		
Item 12: Ä1.3	*Ä1*		**Ä2**	Ä1		
Item 13: Z1.4	*Z1*		Z1	Z1		
Item 14: Z2.5	*Z2*		Z2	Z2		
Item 15: E3.4	*E3*		E3	E3		
Item 16: T2.5	*T2*			T2	T2	
Item 17: E1.4	*E1*			E1	**T2**	
Item 18: T3.3	*T3*			T3	**A3**	
Item 19: E2.1	*E2*			E2	**Z2**	
Item 20: Z3.1	*Z3*			Z3	**A3**	

Rohdaten des Ratings der Items 21-45. R = Raterin bzw. Rater, E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die Nummerierung 1 bis 3 bei den Zuordnungen kennzeichnet die Niveaustufe. Abweichungen von der theoretischen Zuordnung sind grau markiert.

	Theorie	R4	R5	R6	R7	R8	R9
Item 21: Ä3.4	*Ä3*	Ä3	Ä3				
Item 22: T1.1	*T1*	T1	T1				
Item 23: Z2.6	*Z2*	Z2	Z2				
Item 24: A2.2	*A2*	A2	A2				
Item 25: E2.3	*E2*	**Z1**	E2				
Item 26: A3.1	*A3*		**A2**	A3			
Item 27: Z1.3	*Z1*		Z1	Z1			
Item 28: E1.3	*E1*		**E2**	E1			
Item 29: T2.3	*T2*		T2	T2			
Item 30: T1.2	*T1*		T1	T1			
Item 31: E2.4	*E2*			E2	**E1**		
Item 32: A2.3	*A2*			A2	**A1**		
Item 33: Ä1.5	*Ä1*			Ä1	Ä1		
Item 34: Z3.4	*Z3*			Z3	Z3		
Item 35: T3.4	*T3*			T3	T3		
Item 36: Ä3.5	*Ä3*				Ä3	Ä3	
Item 37: Ä2.1	*Ä2*				Ä2	Ä2	
Item 38: E3.3	*E3*				**Z3**	E3	
Item 39: A1.3	*A1*				**A2**	**A2**	
Item 40: Z2.7	*Z2*				Z2	**Z1**	
Item 41: Ä1.6	*Ä1*					Ä1	Ä1
Item 42: Ä2.3	*Ä2*					**Ä1**	Ä2
Item 43: A3.2	*A3*					A3	A3
Item 44: T1.4	*T1*					T1	**Z3**
Item 45: A1.2	*A1*					A1	A1

Anhang 6 – Testheftdesign für das laute Denken.

Testheftdesign für das laute Denken. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die ersten Zahlen kennzeichnen die Niveaustufen. Die zweiten Zahlen stellen die Itemnummer dar, z. B. A2.1 = Alternative Modelle, Niveau 2, Item 1.

Testheft	Items in verwendeter Reihenfolge								
1	A1.1	Ä2.1	Z3.1	Ä1.5	Z2.5	T3.1	Z1.3	T2.3	Ä3.1
2	E1.3	A2.1	E1.4	E3.3	A2.2	Z3.3	T1.2	E2.1	T3.3
3	T1.1	Z2.6	Ä2.3	E3.4	Ä1.3	A2.3	Z1.4	T3.4	Z2.7
4	E1.6	E2.3	E3.6	A1.2	T2.6	A3.1	T1.4	A3.5	Z3.4
5	Z1.2	E2.4	A3.2	A1.3	T2.5	Ä3.4	Ä1.6	Ä2.4	Ä3.5

Anhang 7 – Codierleitfaden zur Auswertung der Denkprotokolle.

Es werden **minimal ganze Sätze** (Codiereinheit) oder – falls dies für die Nachvollziehbarkeit der Zuordnung notwendig ist – **maximal ganze Absätze** (Kontexteinheit) den Codes zugeordnet. **Es werden keine Satzteile, sondern immer ganze Sätze codiert.** Schüleraussagen können auch **mehreren Codes zugeordnet** werden (siehe Beispiele in Kategorien zu den einzelnen Teilkompetenzen, für entsprechende Zuordnung relevanter Teil jeweils unterstrichen).

Nicht-sprachliche Kommentare der SuS wie z. B. Lachen sind in doppelte Klammern gesetzt, für das Verständnis notwendige Ergänzungen bei der Redigierung in eckige Klammern.

Aussagen mit Bezug zu Aufgabenteilen

- **eigene Lösung der Aufgabe**
  - **keine eigene Lösung**

    SuS nennen keine Lösung, die zu der Aufgabe passt, bzw. sagen selbst, dass sie die Aufgabe nicht lösen können.

    Ich kann keine Vermutung ausüben, deswegen muss ich erstmal die Aussagen sehen.

    Mir ist nicht ganz klar, worum es in dieser Aufgabe geht. Für mich hat die Fragestellung mit dem Text nicht viel zu tun, weil das zwei unterschiedliche Sachen sind. Kann ich nicht viel zu sagen.
  - **Beschreibung des Modellobjekts**

    SuS beschreiben das Modellobjekt.

    Man sieht hier einen blauen Luftballon, der in Abschnitte unterteilt ist.

    Es sind immer drei Wände, eine ist unten als Grundlage und dann sind zwei davon an zwei Seiten aufgebaut, die miteinander verbunden sind.
  - **Bewertung des Modells**

    SuS bewerten, wie gelungen das Modell ist. Dabei können sie Kriterien nennen, nach denen sie vorgehen.

    Ich würde sagen, dass ich Marcs Modell besser finde, weil das von Tobias so einheitlich ist und Zellen nicht alle gleich aussehen.

    Das ist zwar jetzt nicht die Aufgabe, aber ich würde sagen, das ist ein gutes Modell.

Was ich bei beiden erst einmal sehe ist, dass die Lunge mit einem Luftballon dargestellt ist, was gut ist, da der sich ausdehnen und auch zusammenziehen kann.

- **Inhaltliche Übereinstimmung von eigener Lösung und Antwortmöglichkeit**

  Die von SuS formulierte Antwort stimmt mit einer der Antwortmöglichkeiten überein. Dies ist auch dann der Fall, wenn nicht der Wortlaut, sondern der Gedanke übereinstimmt. Die Übereinstimmung kann bei der Reflexion der Antwortmöglichkeiten erkannt werden, dies ist für die Einordnung in diesen Code aber nicht erforderlich.

  Das Modell stimmt mit der Form überein, die das doppelte S ist. → „Das Drahtmodell und die Wirbelsäule stimmen in der Form überein."

  Als Vermutung, vielleicht erklärt er [der Aufbau] die Beweglichkeit der Flossen. → „Den Zusammenhang zwischen Aufbau der Flosse und Beweglichkeit"

  Die Geschwindigkeit der Kaubewegung muss bei einem schneller sein und bei einem anderen weniger schnell, und dann muss die Verteilung der Geschmacks- und Geruchsstoffe im Mund gemessen werden. → „Ein (K)Automat „zerkaut" das gleiche Nahrungsmittel einmal schneller, einmal langsamer und die entstehenden Geschmacks- und Geruchsstoffe werden gemessen."

  Ich denke, das eine ist die ganz normale Wirbelsäule, wenn man aufrecht steht, und das andere mit dem Gewicht dran, also das andere Modell mit dem Gewicht dran, zeigt, wenn du eine schwere Last trägst oder wenn du dich irgendwie beanspruchst. Dadurch würde ich erklären, wenn man die Wirbelsäule bewegt oder wenn du ein größeres Gewicht an deinem Körper hängen hast, dass es das dann unterscheidet. → „Welchen Zusammenhang kann Sandra mit dem Drahtmodell der Wirbelsäule erklären? – Den Zusammenhang zwischen der Form der Wirbelsäule und ungesunder Körperhaltung"

- **Argumentationsebene mit Blick auf relevante Teilkompetenz**

  siehe Kategorien zu den einzelnen Teilkompetenzen

- **Bezug zu anderer Teilkompetenz**

  siehe Kategorien zu den einzelnen Teilkompetenzen

- **Bewertung der Antwortmöglichkeiten**
  - **Beschreibung des Modellobjekts**

    siehe oben

  - **Bewertung des Modells**

    siehe oben

  - **Relevanz für Aufgabenstellung**

    SuS begründen die Auswahl einer Antwort damit, dass nicht alle Antworten für die Aufgabenstellung relevant sind. Die Beantwortung ist demnach auf die Zielrichtung der Fragestellung fokussiert.

Man könnte eventuell auch andere Sachen ankreuzen, die aber dann nicht so relevant wären.

Das hat gar nichts mit der Fragestellung zu tun, das kann man gleich durchstreichen.

Der Körperbau – ich denke, das ist das Zutreffendste auf die Fragestellung.

Die ersten drei Antworten sind zwar auch wichtig, aber die beantworten die Frage nicht.

- **Plausibilität der Antwort**

  SuS begründen die Auswahl einer Antwort damit, dass Antworten nicht plausibel sind, ohne dass dies näher erläutert wird.

  Die letzte Möglichkeit mit der anderen Oberfläche, die für einen anderen Körperbau verantwortlich ist, erscheint mir nicht so logisch, weil ich mir das nicht vorstellen kann.

  ((Vorlesen Antwort 1)) Das glaub ich nicht, das schließ ich aus.

  ((schmunzelt)) Das finde ich gut, die Mäuse heilen, so dass das Computermodell wieder funktioniert.

  Ich würde viertens nehmen, also dass sie einmal ein Frostschutzmittel im Blut haben und eine dicke Hornhaut an den Füßen, weil mir das am logischsten vorkommt und weil es irgendwie ((zögert)) auch das realistischste ist. Ich würde vier nehmen.

- **Argumentationsebene mit Blick auf relevante Teilkompetenz**

  siehe Kategorien zu den einzelnen Teilkompetenzen

- **Bezug zu anderer Teilkompetenz**

  siehe Kategorien zu den einzelnen Teilkompetenzen

- **Ja/Nein-Antworten**

  SuS begründen nicht, warum sie sich für oder gegen eine Antwort entscheiden. Sie lehnen eine Antwortmöglichkeit ab oder befürworten sie, ohne sich weiter dazu zu äußern.

  ((Vorlesen Antwort 1)) Nein.

  Das erste ist richtig.

  Ich würde hier b sagen.

- **Expliziter Bezug zu eigener Lösung der Aufgabe**

  SuS begründen die Auswahl einer Antwort damit, dass die selbst die Aufgabe so beantwortet haben. Somit stellen sie einen expliziten Bezug zu ihrer eigenen Lösung her.

  Das erscheint mir am logischsten, das habe ich selbst schon gesagt.

Die erste Antwortmöglichkeit, „es gibt unterschiedliche Modelle, um Tyrannosaurus von außen und von innen zu zeigen", ist ziemlich ähnlich zu dem, was ich davor gesagt hatte. Deswegen würde ich das ankreuzen.

((Vorlesen Antwort 1)) Das hatte ich schon erläutert, warum ich das nicht für richtig halte.

Ich würde sagen, das war das letzte, weil ich es mir so auch selbst erklärt hätte, dass beim Modell 1 eingeatmet wird und bei Modell 2 ausgeatmet.

Aktivierung von Vorwissen

SuS aktivieren zur Beantwortung der Frage Vorwissen, das nicht in der Aufgabe präsentiert wird. Sie können explizit darauf hinweisen, dies muss für die Einordnung in diesen Code aber nicht zwingend der Fall sein.

Wobei ich denke, dass das geschlossene Kreisläufe sind. Das eine war Körper und das andere Lunge oder so.

Man weiß, dass bei höherer Temperatur mehr Algen vorhanden sind. Das heißt, wenn ich ein Becken mit dreißig Grad habe und da schön die Sonne drauf brettert, dann bricht so ein Ökosystem auch zusammen.

Das kann möglicherweise gehen, weil dieses Modell am Ende so einen kleinen Knick hat, da kann man was anhängen, zum Beispiel so einen Federkraftmesser, die wir aus Physik kennen.

Die Kategorisierung innerhalb der einzelnen Teilkompetenzen sowie deren Beschreibung basieren stark auf den Kategorien, die Juliane Grünkorn (Freie Universität Berlin) zur Auswertung von Schülerantworten in offenem Format entwickelt hat.

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Eigenschaften von Modellen **I** Modelle als Kopien *von* etwas	**Maßstabsgetreue Kopie**	Das gezeigte Modell gleicht dem Original bzw. ist mit diesem identisch. Es zeigt das Original evtl. in anderer Größe, aber maßstabsgetreu. SuS können die Notwendigkeit thematisieren, einen Maßstab anzugeben.	*Modelle sind zum Veranschaulichen, also möchte man da eine gewisse Sache besonders gut sehen, zum Beispiel vergrößern oder vielleicht auch nur das Größenverhältnis darstellen.* *(…) Um Tyrannosaurus rex originalgetreu zu zeichnen, müssten sie natürlich die richtigen Schuppenplatten zeichnen.*
	Sehr ähnlich	SuS nehmen das gezeigte Modell als nahezu (ähnlich/fast identisch/ziemlich) originalgetreue Kopie wahr. SuS können Unzufriedenheit mit dem gezeigten Modell bzw. mit dem Modellierungsprozess äußern und damit auf ein nicht gelungenes Duplikat hinweisen. Sie erheben aber den Anspruch, dass ein Modell eine Kopie sein sollte.	*Wenn Insektenaugen gebogen sind, müsste man natürlich das Modell auch in eine gebogene Form bringen, damit es nah an die Wirklichkeit herankommt.* *Es trifft ein Blatt vielleicht annähernd. Ich würde darauf vielleicht noch ein bisschen näher eingehen und nicht so ein wischiwaschi Modell nehmen.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Eigenschaften von Modellen **II** Modelle als idealisierte Repräsentationen *von* etwas	**Keine genaue Nachahmung**	SuS erkennen, dass das Modell keine detailgetreue Nachahmung des Originals ist, führen aber nicht aus, inwiefern Modell und Original sich unterscheiden. Dabei erheben sie nicht den Anspruch, dass das Modell eine Kopie sein sollte.	*Der Luftballon ist nur ein Modell und keine genaue Nachahmung von einem Regenwurm.* *Ein Modell kann nie genau so sein, dass es allem genau entspricht. Das ist auch nur ein Modell, so wie der Durchschnitt ungefähr aussieht.* *Es ist nur ein Modell, natürlich verfälscht es die Ergebnisse.*
	Vereinfachte Darstellung	SuS erkennen, dass das gezeigte Modell eine idealisierte bzw. vereinfachte Darstellung des Originals ist. In dem gezeigten Modell werden nur bestimmte Merkmale oder Eigenschaften hervorgehoben.	*(…) die Tonkörner sind ein bisschen vereinfacht dargestellt, weil man nie so richtig weiß, wie viele da sind.* *Ich würde mir das so erklären, dass die Stoffe nicht alle gleich aussehen, sondern unterschiedliche Formen haben können. Es geht nur darum, dass sie genau zueinander passen und nicht um das genaue Aussehen an sich. Deswegen gibt es wahrscheinlich unterschiedliche Modelle.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Eigenschaften von Modellen **III** Modelle als theoretische Rekonstruktionen *von* etwas –	**Hypothese als Grundlage der Modellierung**	SuS beschreiben, dass die Grundlage eines Modells eine Hypothese ist. Sie können nennen, dass Modelle gezielt entwickelt werden bzw. nicht Tatsachen abbilden.	*Die Grundlage für den Versuchsaufbau ist, dass Fische durch das Wasser schwimmen. Ich denke, das ist die Annahme, so dass die Grundlage dafür ist, dass man den Versuchsaufbau mit Wasser füllt.* *Mir ist eigentlich nicht erklärbar, warum die auf runde Schuppenplatten gekommen sind. Aber das ist wahrscheinlich die häufigste Schuppenplattenform (…).*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Alternative Modelle **Initial** **SuS denken, dass von einem Original nur ein Modell existiert.**	**Verschiedene Originale**	SuS begründen alternative Modelle mit verschiedenen Originalen. Jedes Original ist durch ein Modellobjekt repräsentiert: • unterschiedliche Lebewesen, • unterschiedliche Zellen, • unterschiedliche Menschen u. a.	*Warum es unterschiedliche Modelle gibt – na, weil es unterschiedliche Zellen in den Pflanzen gibt, die sehen nie alle gleich aus (…).* *Ich denke, dass es nicht nur einen Knochenfund gab und auch verschoben in vielen mehreren Kontinenten oder Landteilen und dass er sich dort auch ganz anders anpassen musste und dass man zum Beispiel davon ausging, wenn man einen Knochenfund in kälteren Regionen hat, dass ein dort lebender Tyrannosaurus auch Fell hat, weil das evolutionstechnisch so sein muss, weil man sich immer der Umwelt anpasst, um überleben zu können.*
	Verschiedene Wissensstände über das Original	SuS sind davon überzeugt, dass es zu einem Original nur ein Modell geben kann. Sie können ihre Vorstellung, dass es nur ein Modell zu einem Original geben kann, dadurch beibehalten, indem sie die gezeigten Modelle verschiedenen Zeitpunkten bzw. Erkenntnisständen im Forschungsprozess zuordnen.	*Das ist eben der Lauf der Dinge. Stellt man neue Sachen über eine Sache fest, so stellt man es auch bildlich anders dar. Das heißt, man hat ein erstes Modell, denkt, da fehlt noch das, dann mach ich das doch lieber so. So pusht man sich immer höher, bis man das möglichst bestgetroffene Modell hat.* *Warum es unterschiedliche Modelle gibt – (…) weil es noch nicht hundertprozentig erforscht ist.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Alternative Modelle **I** Unterschiede zwischen den Modellobjekten **SuS denken, dass verschiedene Modelle nur verschiedene Darstellungsvarianten zu einem Original sind.**	**Verschiedene Eigenschaften der Modellobjekte**	SuS begründen alternative Modelle mit verschiedenen Objekteigenschaften. Sie vergleichen die gezeigten Modellobjekte miteinander und beschreiben gezeigte Unterschiede: • verschiedene Darstellungen (z. B. Skizze, 3D), • verschiedene Beschreibungen der gezeigten Modelle (z. B. Bausteine zu sehen, Bausteine nicht zu sehen) u. a. • verschiedene Modelleigenschaften (u. a. beweglich und unbeweglich, weich und hart)	*Letztendlich entspricht das obere Modell mehr oder weniger dem unteren, nur dass verschiedene Materialien benutzt wurden.* *Es gibt unterschiedliche Modelle zu Tyrannosaurus. (…) Eins, was nur mit den Muskeln ist, eins, was sehr bildlich ist, und eins, was eine Skizze ist (…).*
	Unterschiedlich komplexe Modelle	SuS begründen alternative Modelle mit unterschiedlich komplexen Modellen. Sie vergleichen die Komplexität der Modellobjekte miteinander. Dabei können sie beschreiben, dass für verschiedene Adressaten unterschiedlich schwierige Modelle existieren bzw. verschiedene Modelle (übersichtlich bzw. unübersichtlich) existieren.	*Ein Flossenmodell ist ganz einfach gebaut, man hat es einfach gewölbt, ausgeschnitten und auf ein Papier geklebt. Flossenmodell 2 ist bisschen komplizierter aufgebaut.* *Das obere Modell ist eher eine Hobbybastelei und das untere sieht aus wie von irgendwelchen Wissenschaftlern gebaut (…). Es gibt unterschiedliche Modelle zu Tyrannosaurus, weil es unvollständig ist und weil es total verschiedene Modelle sind. Eins, was nur mit den Muskeln ist (…).*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Alternative Modelle **I** Unterschiede zwischen den Modellobjekten **SuS denken, dass verschiedene Modelle nur verschiedene Darstellungsvarianten zu einem Original sind.**	**Verschiedene Einsatzmöglichkeiten**	SuS beziehen sich in ihrer Begründung alternativer Modelle auf deren Zweck. Sie vergleichen die Modellobjekte miteinander und beschreiben deren Nutzen (z. B. Modell zur Veranschaulichung, Modell zum Experimentieren).	*Es gibt unterschiedliche Modelle zu Tyrannosaurus (…). Eins (…) wahrscheinlich zur besseren Veranschaulichung.* *Das obere Modell ist eher eine Hobbybastelei und das untere sieht aus wie von irgendwelchen Wissenschaftlern gebaut, um einen Arm zu simulieren. Während das obere eher aussieht wie so ein Nachmachversuch.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Alternative Modelle **II** Ausgangsobjekt ermöglicht Herstellung verschiedener Modelle von etwas **SuS erkennen, dass das Original aufgrund seiner Vielfältigkeit die Konstruktion verschiedener Modelle ermöglicht.**	**Verschiedene Schwerpunkte**	SuS begründen alternative Modelle mit verschiedenen Blickwinkeln/Perspektiven (z. B. innen und außen) bzw. verschiedenen inhaltlichen Foci auf das Original: • Struktur und Funktion, • verschiedene Zustände des Originals • verschiedene Ausschnitte des Originals u. a.	*Was mir erst einmal auffällt ist, dass das Modell 1 aus der Vogelperspektive ist und das Modell 2 sieht mehr aus wie ein Querschnitt von einer Blüte.* *Ich denke, [diese verschiedenen Modelle haben] mit dem Aufbau und der Innen- und Außenansicht [zu tun].* *Einmal betrachtet sie nur die Füße und einmal betrachtet sie den ganzen Körper. Deswegen würde ich sagen, gibt es zwei Versuche, weil es einmal um die Füße geht und einmal um den ganzen Körper, wie der sich verhält.* *Ich denke, dass es unterschiedliche Modelle gibt, weil es vielleicht darauf ankommt, was man genauer sehen möchte. Wenn man sich zum Beispiel die Staubblätter ganz genau angucken will, die sind relativ klein, macht man das Modell so, dass man die besonders gut sehen kann.* *Ich denke, der Arm ist so komplex, dass ein Modell ziemlich vielfältig sein müsste, um alle Aspekte berücksichtigen zu können.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Alternative Modelle **III** Modelle für verschiedene Hypothesen **SuS erkennen, dass die verschiedenen Modelle unterschiedliche Annahmen im Laufe des Modellierungsprozesses darstellen.**	**Verschiedene Annahmen**	SuS begründen alternative Modelle mit verschiedenen Annahmen/Ideen über das Original. Sie erkennen die Vorläufigkeit von Modellen.	*Dann denke ich darüber nach, ob man irgendeinen Anhaltspunkt zur Farbe oder Fellform in den Knochen sieht, was nicht so ist. Deshalb glaube ich, dass Modell 2 und 3 sich deshalb voneinander unterscheiden, weil offensichtlich kein Stück Haut übrig geblieben ist und man nicht sehen kann, ob der Tyrannosaurus rex Fell hatte oder nicht.* *Meine Antwort würde lauten, es gibt verschiedene Modelle, weil es dazu unterschiedliche Meinungen gibt, wie in der Psychologie bei irgendwelchen Persönlichkeitstheorien, da gibt es auch verschiedene Theorien.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Zweck von Modellen **I** Modellobjekt zur Beschreibung *von* etwas einsetzen **SuS nutzen Modelle zur Beschreibung von Sachverhalten.**	**Darstellung eines Sachverhaltes**	SuS sehen den Zweck von Modellen in der Darstellung des Sachverhaltes und beschreiben, was sie erkennen bzw. was nicht im Modell dargestellt ist. Dabei nennen sie keine Zusammenhänge.	*Das einzige, was ich an diesem Modell sehen kann – ich sehe da überhaupt kein Ufer-Modell, ich sehe da nur Pflanzen, die in der Ecke vom Raum stehen. Das einzige, was ich hieraus sehen könnte, ist, dass an Ufern Pflanzen leben.* *Wie groß die Speiseröhre ist, kann man da [aus dem Speiseröhrenmodell] nicht entnehmen, weil da nichts von wegen 1:1 steht.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Zweck von Modellen **II** Bekannte Zusammenhänge und Korrelationen *von* Variablen im Ausgangsobjekt erklären **SuS nutzen Modelle zur Darstellung von Zusammenhängen.**	**Wiedererkennung von Zusammenhängen**	SuS sehen den Zweck von Modellen im Wiedererkennen von Zusammenhängen im Ausgangsobjekt oder von beobachtbaren Prozessen. Diese werden nicht weiter ausgeführt. Die Modelle werden genutzt, um daran bekannte Tatsachen zu verdeutlichen.	*Ich denke, dass das Schlucken der Nahrung etwas mit der Öffnung der Speiseröhre zu tun hat, aber ich denke auch, dass das Gewicht der Nahrung und die Verformbarkeit der Speiseröhre etwas miteinander zu tun haben.* *Dann würde ich sagen, dass dieses Modell den Zusammenhang zwischen Aufbau der Flossen und Verhalten bei Druck darstellt (…).*
	Erklärung von Zusammenhängen	SuS sehen den Zweck von Modellen im Wiedererkennen und Erklären von Zusammenhängen im Ausgangsobjekt oder von beobachtbaren Prozessen. Sie erkennen diese Zusammenhänge und führen sie aus. Die Modelle werden genutzt, um daran bekannte Tatsachen zu verdeutlichen.	*Ich denke, damit kann man den Weg der Nahrung im Zusammenhang mit den Muskeln und der Kontraktion, der Bewegung der Muskeln, erkennen. Wie sich die Muskeln bewegen, wann sie sich bewegen, wenn das Essen runter geht.* *Ich würde sagen, dass das letzte das richtige ist, weil es von der Form und von dem Gewicht abhängt, weil es auch Unterschiede bei den Fischen gibt, die die gleiche Form haben, und dass die schweren Fische immer etwas langsamer sind.*

Niveau	Codebezeichnung	Beschreibung des Codes	*Ankerbeispiele*
Zweck von Modellen **III** Zusammenhänge von Variablen *für* zukünftige neue Erkenntnisse voraussagen **SuS nutzen Modelle als Instrument zur Erkenntnisgewinnung.**	**Hypothesen über das Modell**	SuS nutzen Modelle als Forschungswerkzeug und machen Hypothesen über das Modell. Dabei muss das Original nicht explizit erwähnt werden, wird aber in der Betrachtung des Experiments mitgedacht.	*Ich würde sagen, dass man diesen (K)Automaten einmal schneller kauen lässt und dann das Gas untersucht. Dann stellt man wahrscheinlich fest, dass die Geruchsstoffe überlagert rauskommen.* *Da steht, es bewegt sich mithilfe der Wasserströmung fort, was für mich bedeutet, dass das Modell langsamer sinkt.*
	Hypothesen über das Original	SuS sehen den Zweck von Modellen als Forschungswerkzeug in der Gewinnung von Erkenntnissen über das Original. Sie nutzen Erkenntnisse über das Modell, um Hypothesen über das Original abzuleiten.	*Er kann wahrscheinlich ableiten, dass beim Fisch die Schwanzflosse nicht nur eine Hautfläche ist, sondern dass dort auch noch bewegliche Teile vorhanden sind und dass diese sich bei Druck eben wölbt, weil ein Fisch das braucht, um zu steuern.* *Jana kann ableiten, dass die Jagdfische, die schnell sein müssen, oder schnelle Fische eher länglich und schmal geschnitten sind und langsame Fische dick sind.*

–

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Testen von Modellen **I** Modellobjekt überprüfen **SuS überprüfen den Modellgegenstand an sich, beziehen sich dabei aber nicht auf das Original.**	**Materialprüfung der Modellobjekte**	SuS überprüfen den Modellgegenstand, indem sie eine Materialprüfung bezüglich Beweglichkeit, Stabilität, Elastizität etc. durchführen. Hierfür überprüfen sie, ob das Modell der Prüfung standhält bzw. ob es dabei beschädigt werden kann.	*Ich denke, dass die Modelle aus Papier bestehen, das heißt, dass Papier vielleicht nicht so toll ist, weil wenn es nass wird, wird es labbrig und ist nicht mehr so stabil.* *Sandra überprüft zum Schluss, ob die Modelle gleich gebogen sind.* *Wenn man ein Modell wieder verbiegt oder so, dann nimmt es sowieso nicht ihre Form wieder an, es sein denn, es ist ein bestimmtes Papier, besteht aus anderen Bestandteilen.* *Wenn man keine Pumpe dabei hat, ist natürlich das Modell schlecht, weil ich glaube, es ist ziemlich schwer, einen Fahrradreifen mit dem Mund aufzupusten.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Testen von Modellen **II** Parallelisieren mit dem Ausgangsobjekt; Modell *von* etwas testen	**Vergleich zwischen Original und Modellobjekt**	SuS überprüfen das Modellobjekt, indem sie die Eigenschaften (Struktur und/oder Funktion) des Originals mit denen des Modells vergleichen.	*Die Zellwand soll der Pappkarton sein und der Zellmembran die Tüte.* *Das Modell und das Blatt stimmen in den Löchern überein, durch die die Luft und damit die zusammenhängende Feuchtigkeit in das Blatt oder in die Schüssel reinkommen kann.* *Sehnen sind als Fäden in Modell 1 auch angedeutet.* *Es gibt Algen, die sind das pflanzliche Plankton. Die werden von tierischem Plankton wie Krebsen gegessen, was hier und hier und hier der Fall ist (Antwort 2, 3 und 4).* *Man kann probieren, kleine Steinchen, die ungefähr im gleichen Verhältnis wie die Moleküle zu den Nährstoffen stehen, da reinfallen zu lassen und zu gucken, ob das nach unten sinkt und auf die andere Seite durchkommt, weil wahrscheinlich durch die Biomembran auch die Nährstoffe in die Zelle kommen.*
SuS vergleichen Modellobjekt mit Original.	**Passung zwischen Original und Modellobjekt**	SuS überprüfen das Modellobjekt, indem sie die Eigenschaften des Originals mit denen des Modells vergleichen und notwendige Übereinstimmungen (Passung) zwischen Modell und Original nennen. Sie geben an, unter welchen Bedingungen das Modell als ein gutes Modell anzusehen ist bzw. welches Modell verwendet werden sollte.	*Dazu wie sich die Speiseröhre in den anderen Organen bewegt, müsste man in dem Modell noch die anderen Organe haben.* *Ich würde sagen, dass das letzte nicht so gut ist. Das ist so wenig und die Fruchtblätter sind nicht so klein.* *Genau, ich bin dafür, dass man mehrere Schläuche verwenden müsste (Antwort 3), weil das bei den Zellen auch so funktioniert, dass es nicht nur eine ist, sondern immer mehrere.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Testen von Modellen **III** Überprüfen von Hypothesen bei der Anwendung; Modell *für* etwas testen **SuS beschreiben, wie eine Hypothese über das Original getestet werden kann.**	**Forschungsdesign**	SuS beschreiben, wie sie eine Vermutung mit dem Modell überprüfen können. Dabei reflektieren sie die Versuchsbedingungen und beschreiben, wie sie mit Blick auf die Aussage, die getroffen werden soll, aussehen müssen.	*Die Nahrungsstücke und die Masse müssen gleich bleiben. Nicht dass der eine zuerst die Nahrungsstücke aufkaut und der andere dann das Aufgekaute nimmt, sondern dass sie beide den gleichen Ausgangspunkt haben.* *Ich denke, dass die Knete nicht aus Eigenantrieb schwimmen kann und es deshalb angemessen ist, sie mit Gewichten zu ziehen, solange die Gewichte alle gleich schwer sind.* *Wie kann Ingo diese Vermutung testen, ganz einfach indem er sich einen möglichst windigen Ort sucht und dieses Modell so anbringt, dass da Wind wirken kann und dann abwartet, was passiert.* *Man kann probieren, kleine Steinchen, die ungefähr im gleichen Verhältnis wie die Moleküle zu den Nährstoffen stehen, da reinfallen zu lassen und zu gucken, ob das nach unten sinkt und auf die andere Seite durchkommt, weil wahrscheinlich durch die Biomembran auch die Nährstoffe in die Zelle kommen.* *Er kann seine Vermutung testen, wenn er in die Mitte von dem jeweiligen Modell drückt (…).* *Ich habe mir gerade darüber Gedanken gemacht, ob das bei solch kleinen Modellen wirklich wahrheitsgetreu ist, weil diese kleinen Formen nicht so viel Wasserwiderstand haben und nicht so viel Wasser verdrängen müssen wie größere Fische, und ob das dann so einfach ist, das mit so einem kleinen Modell zu belegen, mit welcher Form die am schnellsten sind.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Ändern von Modellen **Initial** **SuS denken, dass Modelle nicht verändert werden müssen.**	**Keine Änderung von Modellen**	SuS äußern explizit, dass das gezeigte Modell nicht verändert werden muss. Dabei nennen sie keine Begründung, die einen Bezug zu einer Verbesserung des Modells o. ä. beinhaltet.	*Nee, ich denke dass sie die Modelle mit gleicher Form verwenden sollte, also dass sie nichts ändert.*
Ändern von Modellen **I** Mängel am *Modellobjekt* beheben **SuS sehen die Ursache für Änderungen in Mängeln am Modellobjekt.**	**Verbesserung des Modells**	SuS begründen ein Ändern des Modellgegenstandes mit Möglichkeiten, den Modellgegenstand zu optimieren. Die Begründung zur Optimierung kann z. B. technischer (u. a. bessere Qualität bzw. Leistungsfähigkeit), pädagogischer/didaktischer (u. a. besseres Verständnis), ästhetischer (schöneres Aussehen) Natur sein. SuS begründen, warum ein Modell nicht verändert werden muss, damit, dass es zu keiner Verbesserung des Modells führen würde.	*Hier wurde genannt, dass sie hier feststellt, dass sich der Papierstreifen b schließt. Also muss sie diesen Papierstreifen, also dieses Modell, so verändern, dass dieser Papierstreifen sich nicht schließt.* *Man kann es nicht genau sehen, aber diese Einzelaugen in ihrem Insektenaugenmodell sind rund, weshalb auch da ein Zwischenraum entsteht, als wenn die viereckig wären. (…) Wenn es nicht das Modell total verändern würde, wenn man vielleicht viereckige Einzelaugen anordnet, würde ich das machen.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Ändern von Modellen **II** Modell als Modell *von* etwas durch neue Erkenntnisse oder zusätzliche Perspektiven revidieren **SuS sehen die Ursache für Änderungen im Original.**	**Neue Erkenntnisse über das Original**	SuS begründen Ändern des Modellgegenstandes mit neuen Erkenntnissen über das Original, die beispielsweise durch eine verbesserte Technik gewonnen wurden.	*Stellt man neue Sachen über eine Sache fest, so stellt man es auch bildlich anders dar. Das heißt, man hat ein erstes Modell, denkt, da fehlt noch das, dann mach ich das doch lieber so. So pusht man sich immer höher, bis man das möglichst bestgetroffene Modell hat.* *Im Text stand schon, dass die Wissenschaftler zuvor ein Modell mit runden Schuppen gebaut haben, man aber erkannt hat, dass die versteinerten Schuppen sechseckig sind. Deshalb müssten sie logischerweise die Schuppenform verändern.* *Man muss einfach einen Einfluss hinzufügen, natürlich einen Einfluss, der nur alle drei Jahre berücksichtigt wird.* *Dann würde ich die Möglichkeit auswählen, dass sie am Rücken Haut und am Bauch einen Panzer einzeichnen, weil die das herausgefunden haben und im Prinzip nur das Modell umzeichnen müssen.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Ändern von Modellen **II** Modell als Modell *von* etwas durch neue Erkenntnisse oder zusätzliche Perspektiven revidieren **SuS sehen die Ursache für Änderungen im Original.**	**Mangelnde Passung mit dem Original**	SuS begründen ein Ändern des Modellgegenstandes mit einer mangelnden Passung zwischen Modell und Original. SuS begründen, warum ein Modell <u>nicht</u> verändert werden muss, damit, dass es zu keiner verbesserten Passung von Modell und Original führen würde.	*Oben erscheint mir die Wirbelsäule etwas komisch, ich weiß nicht. Wenn das der Kopf sein soll, eigentlich geht die Wirbelsäule nicht mehr bis in den Kopf, das heißt nur bis in die Nackenregion. Weil der Draht auch nach oben weiter verläuft, weiß nicht, würde ich das vielleicht leicht kürzen.* *Ich denke, dass sie nichts ändern muss, weil augenscheinlich das Drahtmodell mit dem doppelförmigen S, was dem Menschen eher ähneln soll, biegsamer und nicht so hart ist, da es auch so ist.* *An seinem Modell müsste es irgendeine Verbindung zu dieser Verbindung hinten geben, wenn das aufklappt, dass die Zähne auch nach vorne klappen.*

Niveau	Codebezeichnung	Beschreibung des Codes	Ankerbeispiele
Ändern von Modellen **III** Modell *für* etwas aufgrund falsifizierter Hypothesen revidieren **SuS erkennen bei der Anwendung des Modells die Ursache für eine Modelländerung in der fehlenden Passung zwischen Modell- und Originaldaten.**	**Veränderung der Hypothese**	SuS beschreiben, dass die jeweils zugrunde liegende Hypothese sich verändert hat. Sie beschreiben aber nicht, dass sich deshalb auch das Modell ändern muss.	*Regenwürmer haben noch diese kleinen Borsten, die wahrscheinlich wie so eine Art Widerstand sind, womit sie sich nach vorne ziehen. (Hypothese vorher: Regenwürmer bewegen sich dadurch fort, dass sie einzelne Körperabschnitte abwechselnd strecken und zusammenziehen.)* *Ich würde sagen, die Kästchen von dem karierten Papier sind größer als diese wirklichen Einzelaugen. Deswegen nimmt das echte Insektenauge viel mehr ganz kleine scharfe Bilder auf und nicht nur so grobe und deswegen wird das dann im Endeffekt genauer. (Hypothese vorher: Viele große „Einzelaugen" liefern ein sehr scharfes Bild.)*
	Veränderung des Modells wegen Veränderung der Hypothese	SuS begründen Änderungen an Modellen damit, dass die jeweils zugrunde liegende Hypothese sich verändert hat und somit auch das Modell verändert werden muss.	*[hier habe ich keins gefunden, bin mir aber nicht sicher, ob das auch wirklich so ist – ist manchmal schwer von mangelnder Passung abzugrenzen]*

-

Anhang 8 – Übersicht über die Häufigkeiten der Schüleraussagen (*N* = 505). Die Zeilen entsprechen den angesteuerten Teilkompetenzen und Niveaustufen, die Spalten denen, in die die Schüleraussagen codiert wurden.

Items \ Aussagen		Eigenschaften von Modellen			Alternative Modelle				Zweck von Modellen			Testen von Modellen			Ändern von Modellen			
		N I	N II	N III	N 0	N I	N II	N III	N I	N II	N III	N I	N II	N III	N 0	N I	N II	N III
Eigenschaften von Modellen	N I	6	5						2		1	1	17				1	
	N II	2	12						5				9					
	N III			5		1	1		3			1	6				1	
Alternative Modelle	N I	2	2		3	8	4	1	2				2					
	N II	7		1	3	10	5		5				14	1			1	
	N III				2	2	7	3	4		3		14					
Zweck von Modellen	N I	5	1						25				9					
	N II	1				1			13	15	1		9	3				
	N III	3			3	1			4	9	8		10	5				
Testen von Modellen	N I		1						1		1	9	7	3				
	N II												25					
	N III	1							1	1	5	3	11	15		1		
Ändern von Modellen	N I									2			23	1	2	9	8	
	N II	1		1									2				14	
	N III	1	5						1				11	7		14	7	3

Anhang 9 – Psychometrische Qualität der Items bei der empirischen Beschreibung von Modellkompetenz.

Überblick über die psychometrische Qualität der Items zur Dimension Kenntnisse über Modelle. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die zweite Stelle der Itembezeichnung (Nummerierung 1 bis 3) kennzeichnet die Niveaustufe, die dritte die Itemnummer.

Item	n	Schwierigkeit	Distraktorenprüfung			wMNSQ	T	Trenn-schärfe
E1.3	226	-.434	57.1	14.3	28.6	1.00	0.0	.83
E1.4	226	-2.974	50.0	8.3	41.7	0.97	0.0	.47
E2.1	224	-0.821	16.0	28.0	56.0	1.01	0.1	.68
E2.4	224	0.389	41.5	26.2	32.3	1.07	1.5	.67
E3.4	224	-0.645	55.6	28.9	15.6	0.96	-0.7	.74
E3.6	224	-0.431	28.0	12.2	59.8	0.94	-1.1	.75
A1.1	233	-0.694	9.5	59.0	31.4	0.97	-0.6	.62
A1.2	233	-0.176	27.0	32.4	40.5	1.02	0.5	.62
A1.3	233	-1.664	13.5	43.2	43.2	0.95	-0.5	.70
A2.1	228	-0.930	41.0	18.8	40.2	0.99	-0.2	.59
A2.2	228	0.155	25.0	37.5	37.5	1.05	1.3	.53
A2.3	228	0.175	17.5	66.7	15.8	1.08	2.0	.57
A3.1	225	-0.342	18.1	44.8	37.1	0.97	-0.5	.61
A3.2	225	0.137	29.9	21.1	49.0	0.96	-0.9	.48
A3.5	225	0.670	26.9	37.6	35.5	1.10	1.7	.58

Überblick über die psychometrische Qualität der Items zur Dimension ‚Modell bildung'. E = Eigenschaften von Modellen, A = Alternative Modelle, Z = Zweck von Modellen, T = Testen von Modellen, Ä = Ändern von Modellen. Die Nummerierung 1 bis 3 bei den Zuordnungen kennzeichnet die Niveaustufe.

Item	n	Schwierigkeit	Distraktorenprüfung			wMNSQ	T	Trenn-schärfe
Z1.2	225	-1.896	32.1	25.0	42.9	1.02	0.2	.55
Z1.4	225	-0.594	37.7	22.1	40.3	0.99	-0.1	.73
Z2.5	227	0.585	34.8	27.5	37.7	1.07	1.3	.66
Z2.6	227	-0.769	42.0	55.7	2.3	0.95	-0.7	.61
Z2.7	227	-0.329	30.1	15.1	54.8	0.95	-1.0	.63
Z3.1	227	1.090	41.4	27.8	30.8	1.04	-0.3	.70
Z3.3	227	0.336	31.6	34.2	34.2	1.02	0.1	.57
Z3.4	227	-0.724	18.2	11.3	70.4	0.99	-0.2	.51
T1.1	228	-0.663	46.1	9.2	44.7	0.99	-0.1	.61
T1.2	228	0.173	75.2	9.1	15.7	1.04	1.0	.55
T1.4	228	0.193	12.3	39.3	48.4	0.99	-0.3	.51
T2.3	233	-1.111	42.6	39.3	18.0	1.01	0.1	.60
T2.5	233	-0.334	26.2	37.9	35.9	0.96	-1.0	.67
T2.6	233	-1.281	56.1	33.3	10.5	1.03	0.4	.49
T3.1	225	-1.098	41.3	44.4	14.3	0.98	-0.3	.55
T3.3	225	0.445	42.6	50.4	7.1	1.00	0.0	.63
T3.4	225	-0.583	62.9	19.1	18.0	1.02	-0.4	.60
Ä1.5	230	-0.261	38.6	26.3	35.1	0.99	-0.3	.66
Ä1.6	230	-1.186	43.4	30.3	26.3	0.96	-0.5	.73
Ä2.1	228	0.036	44.6	32.2	23.1	0.95	-1.1	.71
Ä2.3	228	-0.340	13.4	58.8	27.8	1.01	0.2	.68
Ä2.4	228	-1.302	34.5	21.8	43.6	0.89	-1.4	.68
Ä3.1	225	1.182	14.8	30.8	54.4	1.09	1.2	.46
Ä3.4	225	-0.379	13.2	47.1	39.7	1.02	0.5	.64
Ä3.5	225	0.118	33.3	49.0	17.7	1.03	0.7	.57

Anhang 10 – Korrelationen zwischen Modellkompetenz und anderen Variablen.

Korrelationen zwischen Modellkompetenz und Geschlecht, allgemeinen kognitiven Fähigkeiten, Lesegeschwindigkeit, Leseverstehen, Wissenschaftsverständnis sowie Chemieunterricht.

	Geschlecht	allgemeine kognitive Fähigkeiten	Lesegeschwindigkeit	Leseverstehen	NOS	kein Chemieunterricht
PV1_imp1	.025	.377**	.546**	.504**	-.082**	.198**
PV2_imp1	-.001	.374**	.544**	.500**	-.077**	.213**
PV3_imp1	.036	.336**	.531**	.484**	-.072*	.224**
PV4_imp1	.008	.366**	.510**	.463**	-.064*	.235**
PV5_imp1	.000	.376**	.522**	.477**	-.070*	.183**
PV1_imp2	.007	.365**	.558**	.514**	-.059*	.202**
PV2_imp2	-.020	.339**	.524**	.479**	-.059*	.201**
PV3_imp2	.018	.360**	.532**	.485**	-.066*	.239**
PV4_imp2	.007	.346**	.535**	.489**	-.061*	.226**
PV5_imp2	.002	.361**	.526**	.481**	-.078**	.228**
PV1_imp3	.020	.382**	.536**	.493**	-.054	.196**
PV2_imp3	.012	.358**	.530**	.488**	-.050	.195**
PV3_imp3	.016	.377**	.545**	.503**	-.069*	.230**
PV4_imp3	.007	.332**	.548**	.503**	-.031	.234**
PV5_imp3	.012	.354**	.528**	.483**	-.067*	.215**
PV1_imp4	.026	.384**	.523**	.481**	-.048	.174**
PV2_imp4	.009	.374**	.544**	.497**	-.044	.206**
PV3_imp4	.017	.352**	.158**	.475**	-.081**	.246**
PV4_imp4	-.034	.363**	.538**	.494**	-.026	.245**
PV5_imp4	.024	.355**	.528**	.482**	-.068*	.195**
PV1_imp5	.014	.359**	.523**	.480**	-.014	.192**
PV2_imp5	-.005	.315**	.524**	.481**	-.023	.207**
PV3_imp5	.026	.335**	.528**	.481**	-.010	.240**
PV4_imp5	.017	.332**	.539**	.497**	.004	.232**
PV5_imp5	.009	.330**	.524**	.477**	.004	.229**

Fortsetzung Tab.

	Geschlecht	allgemeine kognitive Fähigkeiten	Lesegeschwindigkeit	Leseverstehen	NOS	kein Chemieunterricht
PV1_imp6	.011	.342**	.534**	.492**	-.064*	.186**
PV2_imp6	-.004	.325**	.527**	.478**	-.075*	.202**
PV3_imp6	.009	.332**	.518**	.474**	-.052	.229**
PV4_imp6	-.008	.345**	.528**	.482**	-.560	.225**
PV5_imp6	.011	.354**	.515**	.474**	-.057	.216**
PV1_imp7	.008	.344**	.538**	.500**	-.106**	.186**
PV2_imp7	.012	.318**	.544**	.497**	-.091**	.219**
PV3_imp7	.013	.331**	.521**	.475**	-.072*	.232**
PV4_imp7	.008	.328**	.531**	.481**	-.086**	.226**
PV5_imp7	.037	.331**	.532**	.484**	-.089**	.233**
PV1_imp8	.011	.379**	.530**	.490**	-.071*	.187**
PV2_imp8	-.017	.339**	.529**	.478**	-.048	.212**
PV3_imp8	.023	.345**	.509**	.467**	-.058	.218**
PV4_imp8	.002	.339**	.521**	.478**	-.084**	.227**
PV5_imp8	.019	.364**	.522**	.475**	-.080**	.214**
PV1_imp9	.030	.395**	.539**	.497**	-.057	.190**
PV2_imp9	-.013	.352**	.537**	.493**	-.069*	.191**
PV3_imp9	.021	.353**	.527**	.484**	-.089**	.228**
PV4_imp9	-.018	.360**	.523**	.478**	-.063*	.229**
PV5_imp9	-.002	.381**	.517**	.472**	-.070*	.219**
PV1_imp10	.304	.346**	.553**	.508**	-.061*	.186**
PV2_imp10	-.004	.328**	.535**	.487**	-.078**	.217**
PV3_imp10	.025	.330**	.533**	.491**	-.090**	.228**
PV4_imp10	.030	.344**	.526**	.484**	-.074*	.231**
PV5_imp10	.017	.346**	.531**	.489**	-.071*	.215**
M	.016	.351	.523	.486	-.072	.215
SD	.0039	.0004	.0059	.0001	.0115	.0003

keine Markierung = nicht signifikant, * = signifikant ($p < .05$), ** = sehr signifikant ($p < .01$).

Korrelationen zwischen Modellkompetenz und Biologie-, Chemie-, Physik-, Mathematik-, Deutschnote sowie Note in der ersten Fremdsprache.

	Biologie-note	Chemie-note	Physik-note	Mathe-matiknote	Deutsch-note	Note erste Fremdsprache
PV1_imp1	-.295**	-.242**	-.344**	-.318**	-.329**	-.239**
PV2_imp1	-.271**	-.264**	-.379**	-.354**	-.350**	-.266**
PV3_imp1	-.278**	-.279**	-.379**	-.303**	-.320**	-.248**
PV4_imp1	-.294**	-.281**	-.398**	-.340**	-.310**	-.244**
PV5_imp1	-.286**	-.221**	-.366**	-.321**	-.311**	-.233**
PV1_imp2	-.313**	-.232**	-.362**	-.329**	-.301**	-.227**
PV2_imp2	-.267**	-.265**	-.376**	-.324**	-.294**	-.264**
PV3_imp2	-.286**	-.256**	-.360**	-.304**	-.301**	-.227**
PV4_imp2	-.319**	-.278**	-.402**	-.334**	-.280**	-.252**
PV5_imp2	-.277**	-.236**	-.357**	-.306**	-.281**	-.236**
PV1_imp3	-.288**	-.259**	-.366**	-.319**	-.275**	-.196**
PV2_imp3	-.263**	-.251**	-.384**	-.337**	-.338**	-.254**
PV3_imp3	-.281**	-.249**	-.382**	-.311**	-.317**	-.218**
PV4_imp3	-.299**	-.286**	-.399**	-.327**	-.286**	-.201**
PV5_imp3	-.270**	-.219**	-.364**	-.315**	-.306**	-.223**
PV1_imp4	-.290**	-.258**	-.342**	-.319**	-.320**	-.215**
PV2_imp4	-.268**	-.270**	-.379**	-.340**	-.323**	-.248**
PV3_imp4	-.292**	-.291**	-.370**	-.301**	-.324**	-.221**
PV4_imp4	-.305**	-.282**	-.400**	-.321**	-.284**	-.202**
PV5_imp4	-.297**	-.214**	-.354**	-.304**	-.321**	-.234**
PV1_imp5	-.307**	-.232**	-.385**	-.335**	-.296**	-.234**
PV2_imp5	-.269**	-.253**	-.403**	-.344**	-.316**	-.266**
PV3_imp5	-.276**	-.266**	-.363**	-.294**	-.303**	-.229**
PV4_imp5	-.297**	-.246**	-.403**	-.325**	-.283**	-.230**
PV5_imp5	-.279**	-.230**	-.364**	-.319**	-.290**	-.240**
PV1_imp6	-.290**	-.231**	-.344**	-.314**	-.305**	-.221**
PV2_imp6	-.282**	-.266**	-.381**	-.339**	-.305**	-.256**
PV3_imp6	-.277**	-.272**	-.365**	-.300**	-.308**	-.242**
PV4_imp6	-.297**	-.262**	-.398**	-.316**	-.260**	-.219**
PV5_imp6	-.293**	-.243**	-.359**	-.313**	-.314**	-.235**

Fortsetzung Tab.

	Biologie-note	Chemie-note	Physik-note	Mathe-matikno-te	Deutsch-note	Note erste Fremdsprache
PV1_imp7	-.276**	-.226**	-.338**	-.296**	-.283**	-.210**
PV2_imp7	-.264**	-.265**	-.351**	-.320**	-.298**	-.235**
PV3_imp7	-.296**	-.254**	-.370**	-.304**	-.310**	-.223**
PV4_imp7	-.280**	-.263**	-.370**	-.319**	-.247**	-.188**
PV5_imp7	-.288**	-.247**	-.331**	-.303**	-.279**	-.222**
PV1_imp8	-.278**	-.207**	-.321**	-.295**	-.313**	-.231**
PV2_imp8	-.260**	-.279**	-.338**	-.325**	-.325**	-.260**
PV3_imp8	-.280**	-.277**	-.363**	-.295**	-.317**	-.264**
PV4_imp8	-.297**	-.259**	-.374**	-.325**	-.301**	-.227**
PV5_imp8	-.287**	-.247**	-.343**	-.308**	-.311**	-.222**
PV1_imp9	-.298**	-.245**	-.340**	-.300**	-.305**	-.213**
PV2_imp9	-.285**	-.256**	-.391**	-.346**	-.326**	-.277**
PV3_imp9	-.290**	-.264**	-.369**	-.294**	-.321**	-.230**
PV4_imp9	-.303**	-.270**	-.395**	-.314**	-.280**	-.219**
PV5_imp9	-.300**	-.237**	-.351**	-.326**	-.304**	-.237**
PV1_imp10	-.308**	-.226**	-.350**	-.318**	-.304**	-.217**
PV2_imp10	-.287**	-.250**	-.390**	-.343**	-.315**	-.243**
PV3_imp10	-.324**	-.247**	-.384**	-.321**	-.325**	-.251**
PV4_imp10	-.311**	-.260**	-.401**	-.328**	-.271**	-.218**
PV5_imp10	-.336**	-.253**	-.374**	-.338**	-.313**	-.241**
M	-.289	-.253	-.369	-.319	-.304	-.233
SD	.0003	.0004	.0004	.0002	.0004	.0004

** = sehr signifikant ($p < .01$).